Humboldt-Universität zu Berlin

DISSERTATION

Lay Internet Usage - An Empirical Study with Implications for Electronic Commerce and Public Policy

Zur Erlangung des akademischen Grades
doctor rerum politicarum
im Fach Wirtschaftswissenschaften

Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin

von
Herrn Diplom-Kaufmann Mario Christ
geboren am 12. Juli 1972 in Berlin

Dekan der Wirtschaftswissenschaftlichen Fakultät: Prof. Michael Burda, Ph.D.

Gutachter: 1. Prof. Oliver Günther, Ph.D.

2. Prof. Ramayya Krishnan, Ph.D.

eingereicht: 04.12.2002
Datum der Promotion: 20.12.2002

Abstract

Despite the substantial social and economic implications of the World Wide Web, there is still a surprising lack of empirical research on Web usage. Specifically, at the level of the *individual* user, little is known about key issues of Internet usage, such as the trajectory of change over time in the number of visits to Web sites, the degree of individual loyalty to Web sites, and the demographics that determine Web usage.

In order to overcome this lack of research, we report in this dissertation the results of several interrelated studies of individual Web usage patterns of average citizens from the Pittsburgh area.

This dissertation advances the research on individual Web usage by:

- analyzing the impact of increasing Web site visiting opportunities on Web utilization rates of individual users,
- employing session-based measures to data on individual Web usage in order to identify how Web users change the way they use the Web as their level of expertise increases,
- analyzing whether different user groups also differ in loyalty to Web sites and whether users converge over time to a set of 'favorite' Web sites,
- specifically dealing with the issue of Web portal utilization to answer the question whether Web portal users are different from average Web users.

We develop measures of Web usage that are particularly relevant from a business and public policy perspective. By applying these measures to longitudinal data on Web usage, we identify significant trends in individual Internet usage. For example, we reveal that individual Web usage is not distributed equally across subgroups of users. Web users can be clustered into four groups with distinct trajectories of Web usage. All groups reach saturation in their extent of Web usage after following a downward path. Further, most Web users spent only limited time in the Web and only a small group of users uses the Web heavily. Also, users show consistently little loyalty to Web sites. Surprisingly, as Web users gain experience in using the Web, there does not seem to be a significant shift from undirected browsing to directed access of Web sites over time.

We apply regression models in order to predict the determinants of Web utilization. Individual characteristics, such as ethnic background, gender, household income, phone usage, e-mail usage, and computer skill level, determine Web usage.

Thus, the results have implications for both electronic commerce and public policy as it pertains to the digital divide. They are particularly useful for marketing departments, especially in the information and communication industry. Discussions of Web user loyalty and Web visiting opportunities as conducted in this dissertation are relevant to business models in use in business-to-consumer electronic commerce, especially for Internet companies that rely on advertising income generated from serving banner advertisements and companies that need to maintain a high degree of customer loyalty. The results also provide the factual foundation for key policy initiatives to promote access to the Internet for all groups of people. Policy makers need data on Internet usage in order to measure the size of a possible digital divide and ensure that everybody belonging to the present and the next generation – and not a subgroup of people only – has access to the Internet.

In summary, this study advances the empirical foundation for understanding individual Web use. The findings of this dissertation will be useful to stakeholders in the new Information Age, in particular marketing departments and policy makers.

Keywords:

User Behavior, Internet and the World Wide Web, Statistical Analysis, Web Portals, Digital Divide

Zusammenfassung

Trotz substantieller ökonomischer und sozialer Implikationen des World Wide Webs existiert noch immer eine überraschend große Forschungslücke in Bezug auf empirische Untersuchungen der Webnutzung. Insbesondere bezüglich der *individuellen* Webnutzung weiß man heute noch wenig über Schlüsselthemen dieses Forschungsfeldes, wie zum Beispiel die Anzahl der Webseitenbesuche von Individuen, der Loyalität von Nutzern, und den demographischen Charakteristika, die bestimmend für die Internetnutzung sind.

Deshalb sieht sich diese Dissertation als Schritt zur Überbrückung dieser Forschungslücke. Sie präsentiert die gewonnenen Erkenntnisse verschiedener, voneinander abhängiger, empirischer Studien der individuellen Webnutzung Pittsburgher Bürger.

Diese Dissertation erweitert die Forschung im Bereich individuellen Webnutzungsverhaltens durch:

- die Analyse des Einflusses der steigenden Anzahl von Webangeboten auf die individuelle Webnutzung,
- die anwendung sessionbasierter Maße auf individuelle Webnutzungsdaten, um Einsichten in den Verlauf der Webnutzung bei gleichzeitigem Anstieg der individuellen Weberfahrung zu erhalten,
- die Analyse der Loyalitaet im Web von einzelnen Nutzergruppen, um die Frage zu beantworten, ob Nutzergruppen zu favorisierten Seiten im Web konvergieren,
- spezifisches Herangehen an das Thema der Portalnutzung im Web und das Beantworten der Frage, ob sich Portalnutzer von durchschnittlichen Internetnutzern unterscheiden.

Aus betriebswirtschaftlicher und volkswirtschaftlicher Sicht interessante Webnutzungsmaße werden entwickelt und diskutiert. Die Anwendung dieser Maße führt zu Erkenntnissen bezüglich signifikanter Trends. So wird beispielsweise deutlich, dass keinesfalls eine Gleichverteilung der Nutzung über Nutzer und Zeit besteht. Nutzer können in vier Gruppen mit verschiedenen Entwicklungskurven eingeteilt werden. Alle Nutzergruppen nähern sich über die Zeit Sättigungsgrenzen der Webnutzung an. Außerdem verbringen die meisten Nutzer nur wenig Zeit im Internet. Auch wird deutlich dass Loyalität im Web äußerst gering ist und Webnutzer trotz steigender Erfahrung im Umgang mit dem Internet nicht sonderlich gezieltes Surfverhalten entwickeln.

Zusätzlich führt die Anwendung von Regressionsmodellen zu Erkenntnissen über die individuellen Charakteristika, welche die Webnutzung beeinflussen. Solch Charakteristika

sind zum Beispiel ethnische Herkunft, Geschlecht, Haushaltseinkommen, Telefon- und Emailnutzung und Computerkenntnisse.

Daher liefert die vorliegende Arbeit Erkenntnisse, welche sowohl aus betriebswirtschaftlicher Sicht als auch aus volkswirtschaftlicher Sicht Relevanz haben. Insbesondere können Marketingabteilungen, vor allem in der Informations- und Kommunikationsindustrie, von den vorliegen Resultaten profitieren. Themen wie Webloyalität und Webnutzung, die in der vorliegenden Arbeit angesprochen werden, sind insbesondere relevant für Geschäftsmodelle aus dem B2C Bereich. Adressaten sind dementsprechend zum Beispiel Internetfirmen, welche von Werbeeinkünften aus Bannerwerbung abhängig sind und Firmen, welche einen hohen Grad an Loyalität unter Ihren Webnutzern suchen. Außerdem bilden die Erkenntnisse die Grundlage für staatliche Initiativen, die der Sicherstellung des Zugangs zum Internet alle Gruppen der Bevölkerung dienen.

Die vorliegende Arbeit reichert die empirische Grundlage, welche zum Verständnis individueller Webnutzung nötig ist, an. Die Erkenntnisse sind insbesondere für am neuen Informationszeitalter teilhabenden Individuen und Institutionen, auch staatlicher Art, interessant.

Schlagwörter:

Nutzerverhalten, Internet und das World Wide Web, Statistische Analyse, Web Loyalität, Web Portale, Digital Spaltung

Table of Contents

Figures

Tables

Abkürzungsverzeichnis

ADO	ActiveX Data Objects
ASP	Active Server Pages
B2B	Business-to-Business
B2C	Business-to-Consumer
B2E	Business-to-Employee
BIC	Bayesian Information Criterion
CGI	Common Gateway Interface
CLF	Common Logfile Format
CPM	Cost per Thousand Impressions
CRM	Customer Relationship Management
DBMS	Database Management Systems
DNS	Domain Name Service
DSL	Digital Subscriber Line
eCRM	Electronic Costumer Relationship Management
GUI	Graphical User Interface
HCI	Human Computer Interaction
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
ISDN	Integrated Services Digital Network
IT	Information Technology
IuKDG	Informations- und Kommunikationsdienstegesetz
JDBC	Java Database Connectivity
JMM	Jupiter Media Metrix

JPEG	Joint Photographics Experts Group
JSP	Java Server Pages
ODBC	Open Database Connectivity
OECD	Organization for Economic Cooperation and Development
TCP	Transmission Control Protocol
UMTS	Universal Mobile Telecommunication Service
URL	Uniform Resource Locator
WWW	World Wide Web

1 Introduction and Motivation

In 2002, over 147 million hosts are connected to the Internet [InternetReport] and 544 million people [NUA] have access to billions of pages of content. While originally conceived for the military and primarily used in academia, the Web turned into one of the most important tools of communication for other sectors. The Internet has already changed how people live. In this regard, the Internet is clearly one of the most important communication innovations in history, and its social and economical implications are substantial. Figure 1 shows the explosive growth in number of Web sites available to users over the period 1995 to 2000.

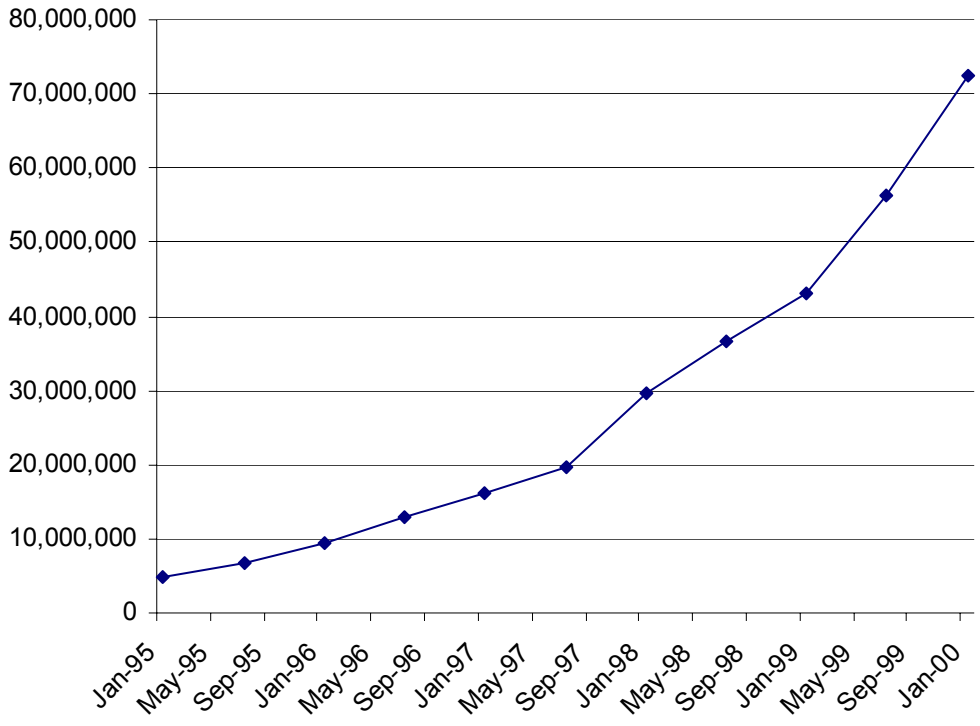


Figure 1: Number of hosts advertised in the DNS [IDS00]

The rapid growth depicted in Figure 1 has led to many new research questions. Arising research issues include the analysis of individual Web usage patterns and the identification of characteristics of individuals that determine Web usage.

There are numerous studies on aggregate Web utilization. They report that the number of sites in the Internet is increasing as well as the number of users [IDS00, NUA]. However,

aggregate utilization always reflects a combination of at least two distinct usage components, such as the number of users and the intensity of their use.

Yet at the level of the *individual* user, little is known about key issues of Internet usage, such as the trajectory of change over time in the number of visits to Web sites, the degree of individual loyalty to Web sites, and the demographics that determine Internet usage. Thus, this dissertation examines how *individuals* responded to the exponential growth in Web site availability.

The Internet may be used from various access points. However, there is a specific lack of research with respect to residential use of the Internet [Choo00]. Therefore, this dissertation aims at overcoming this lack of research on lay Internet usage by:

- developing the right measures of individual Internet use at home,
- extracting the right source of data with these measures,
- describing how average citizens use the Internet,
- identifying significant trends in individual Internet usage,
- predicting the determinants of such usage,
- and statistically analyzing and interpreting the results.

It thereby advances the research on residential Internet usage in many different aspects.

There are several areas that may benefit from an improved understanding of individual Internet usage in general and Web browsing activities in particular:

- The necessity for educators, developers, designers, and managers to understand how people use technology in the context of their daily lives is generally accepted.
- Also, many companies already took the first step to the online world [Bensberg01]. For these companies, it is important to know who is online and which activities are done online. Such online customers can use the Web for online shopping and customer service may interact with the customer. In this regard, conducting such a study on Internet usage and the determining demographic characteristics of usage is necessary for marketing departments, particularly in the information and communication industry.
- In the online world, visits to given Web sites are considered an important measure of market share and success. Therefore, discussions of Web site visiting opportunities as conducted in this dissertation are relevant to business models in use in business-to-consumer electronic commerce, especially for Internet companies that rely on advertising income generated from serving banner advertisements.

- Also, research on Web usage provides the factual foundation for key policy initiatives to promote access to the Internet for all groups of people. Policy makers need data on Internet usage to measure the size of a possible digital divide and ensure that everybody belonging to the present and the next generation – and not a subgroup of people only – has access to the Internet

In summary, the findings of this dissertation will be useful to stakeholders in the new Information Age, in particular policy makers. This study advances the empirical foundation for understanding individual Web use. It shows both expected and unexpected results. Because the results have important implications for electronic commerce and public policy pertaining to the digital divide, each chapter deals specifically with the implication of the results from a business and policy perspective.

The remainder of this dissertation is organized as follows:

The data source used for this dissertation is described in **chapter 2**: “Data Source: The HomeNet Project at Carnegie Mellon University” (pp. 19 ff.). This chapter describes the characteristics of the data sample, deals with data collection and data extraction issues, describes normalizations and assumptions, and evaluates the quality and representativeness of the data sample.

Chapter 3: “Saturation of Lay Web Usage” (pp. 40 ff.) explores whether the increase in Web site visiting opportunities spurred an increase in the Web utilization rates of individual users. It reveals that individual Web usage is subject to saturation by applying a semi-parametric, group-based statistical method. It analyzes the number of distinctive Web sites accessed per month as a measure of the user’s interest in the World Wide Web. It thereby not only finds user groups with different levels of usage, but also identifies distinctive trajectories of the development of Web usage over time and provides demographic profiles of the identified user groups. The resulting trajectories are compared to the overall trend in the number of Web sites, which multiplied exponentially during the period of observation.

In **chapter 4**: “Analyzing Web Sessions” (pp. 62 ff.), time-based measures are applied to the HomeNet data in order to advance the research from chapter 3. Specifically, this chapter aims at detecting changes in Web usage associated with increased experience of using the Web. It uses advanced measures that are based on user sessions in the Web to answer the question whether or not users shift from undirected browsing in the Web to directed access of Web sites as they gain expertise in using the Web. Five key measures of Web usage are computed for each of the subgroups of users identified in chapter 3.

Formal regression models are applied to the HomeNet data in order to identify determinants of session-based Web usage.

Because the success of many websites is inextricably linked to their ability to maintain a high degree of customer loyalty, **chapter 5**: “Web User Loyalty and Web Site Stickiness” (pp. 80 ff.) introduces precise ways of measuring loyalty on the Web and characterizes loyalty empirically using the HomeNet data. This chapter is based on the findings from chapters 3 and 4 that emphasize that research on the dynamics of usage has to incorporate an analysis of churn in the Web. Thus, this chapter addresses the question whether the groups identified in chapters 3 and 4 also differ in loyalty to Web sites. Also, it answers the question whether users converge over time to a set of ‘favorite’ Web sites. Other key measures that are developed in this chapter include popularity of Web sites and Web site stickiness. Popularity influences the probability of a given Web site to be in a given user’s set of favorite sites. Stickiness determines the ability of Web sites to actually remain in this set of favorite domains over time.

Chapter 6: “Portal Utilization” (pp. 99 ff.) is based upon the finding from chapter 5 that Web portals are different from other Web sites in the sense that they achieve high degrees of popularity and stickiness among Web users. Therefore, this chapter addresses the issue of Web portal utilization. It answers the question whether Web portal users are different from average Web users. Specifically, because Web portals need to know who their customers are, this analysis addresses the question what demographic characteristics distinguish Web users who actually make use of additional features of Web portal sites such as yahoo.com on the one hand and Web users who only use the search capabilities of such portals on the other hand. Web portal utilization of individuals is measured and demographic profiles of user groups with different portal utilization levels are developed. A formal regression model is applied to identify demographic characteristics of individuals that determine portal utilization.

Many of the results presented in this dissertation have implications for public policy as it pertains to the digital divide. Therefore, **chapter 7**: “The Digital Divide Exists” (pp. 111 ff.) provides the foundation for a deeper discussion of the digital divide. It presents empirical results in as far as they relate to the digital divide issue and summarizes the key policy issues and the relevant literature on the digital divide of society. Because a digital divide can be put into a global or national context, chapter 7 deals with the issue of the digital divide from an American, a European, and a global perspective.

Finally, **chapter 8**: “Concluding Remarks” (pp. 133 ff.) concludes this dissertation by summarizing the key results.

2 Data Source: The HomeNet Project at Carnegie Mellon University

2.1 Introduction

Services offered online are usually based on surprisingly little information about the preferences of Web users. Despite the Web's growing popularity, there are few direct, rigorous studies of Web browsing behavior [Choo00], which help companies and public institutions in their decisions to provide on-line services. One possible reason for that is the difficulty in collecting complete data sets to describe Web browsing sessions. Existing collections of information on Web use are often based on use in businesses and universities. Moreover, many studies over-represent the higher income, well-educated young males. In this regard, Kraut et al. [Kraut96b] point out that knowledge on Internet usage is based upon the behavior of the predominately upper income, white male professionals. Other studies, such as [McKenzie01a] rely on data on Web usage of university students. In summary, there is a lack of research on residential use of the Web. Thus, Kraut et al. conclude that there is a need for studies in residential settings with average citizens, “to understand how to build and deploy on-line services valuable to households” [HomeNet95]. Thus, the HomeNet project at Carnegie Mellon University[1] was started in 1995. Through detailed monitoring of individual Internet use, periodic surveys, and interviews with family members, the HomeNet project measured the demand for and impact of electronic communication and telecommunication services over time.

This study is based on individual use records from the HomeNet project. HomeNet is a field trial whose aim was to overcome economical and technological barriers to Web usage and understand usage of the Internet at home by lay users. Starting in 1995, it provided families in the Pittsburgh area with hardware and Internet connections for no charge. Other measures to reduce barriers of getting online included customization of the user interface (based on user interests), training sessions on the day the users picked up their computers, and a help desk. Residential usage of on-line services such as electronic mail, computerized bulletin boards, chat groups, and the World Wide Web was carefully documented [Kraut96a, HomeNet].

[1] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.

HomeNet is distinctive in many different aspects. It was explicitly designed to empirically study the human issues around Internet Usage in order to answer two broad questions [Homenet95]:

- Once the Internet is easily available, how will average citizens use it?
- What impact will this usage have on the average citizen's lives?

In this dissertation, we use the data from the HomeNet project to get answers to the first of these two questions. The impact of the Internet on average citizens' lives is beyond the scope of this dissertation and is discussed in the psychological and Human Computer Interaction literature [Kraut96a]. Specifically, studies within the HomeNet project investigated how people's motivations for use of the Internet and its impact on their lives changes over time. They did so by combining automated measures, longitudinal surveys, and home interviews.

A major finding of the HomeNet project is that demographic factors – generation, race, and gender – rather than socioeconomic factors – income and education – or psychological factors – like social extraversion and attitudes towards computing – were the major factors that influence Internet use. The fact that neither household income nor education predict Internet use, does strongly suggest that if economic barriers to Internet access were removed, people across socioeconomic lines would use the Internet. However, gender, race, and generation were all strong predictors of Internet use in the sample. For example, teenagers turned out to be much heavier users than their parents, and among teenagers, boys were heavier users than girls. Another major finding is that Internet use is strong in the aggregate, but varies widely and declines with time [Kraut96b].

These key findings are used as a basis for the research described in subsequent chapters. In this regard, we build upon these findings and advance the research on residential Internet usage by developing and applying sophisticated measures of Internet usage, which are relevant from a business and public policy perspective.

This chapter is organized as follows: Section 2.2 describes the characteristics of the data sample. Section 2.3 deals with data collection issues. Data extraction issues, particularly issues arising from Web log analysis, are discussed in section 2.4. Section 2.5 deals with normalizations and assumptions. Finally, section 2.6 discusses the quality and representativeness of the HomeNet sample.

2.2 Sample Characteristics

This dissertation reports the results of several interrelated studies of Web usage of 139 HomeNet users. It combines data on characteristics of these HomeNet users with results from longitudinal Web log analysis of the URLs accessed during 33916 user-days of Web usage. Individuals used the Internet at home from 11-6-1995 to 4-28-1997. This is just the period of time in which the Web became popular, many people accessed the Internet for the first time, the number of Web services offered to people grew exponentially, and many search engine sites made the transition to Web portal sites.[2] Individuals start their Web use on different starting dates and exhibit different durations of Web usage. On average, Web usage behavior of users was observed for 311 days.

2.2.1 Families

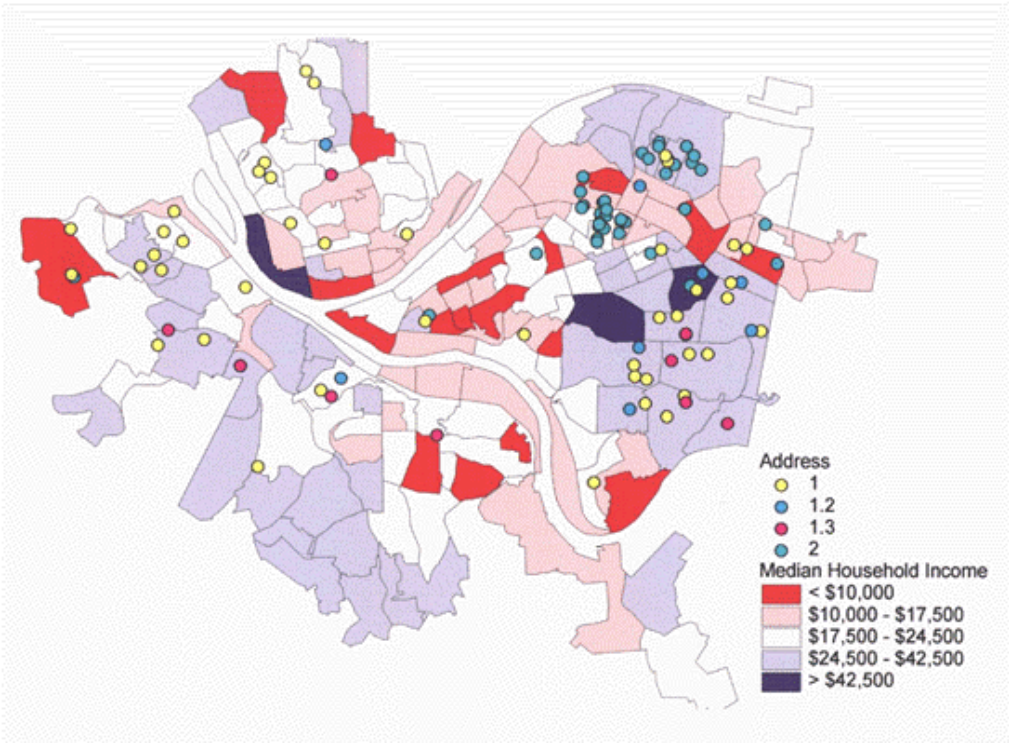


Figure 2: Map of Pittsburgh neighborhood income levels with location of HomeNet participants [HNMAP]

[2] These are welcome characteristics of the period of observation, which is discussed in section 2.5.

The HomeNet data set used for this dissertation consists of data from 139 family members in 56 households.[3] The families were recruited from four public high schools in the city of Pittsburgh, Pennsylvania, USA. Reaching demographic diversity was of utmost importance when selecting the users. Figure 2 shows a map of Pittsburgh neighborhood income levels with location of HomeNet participants.

2.2.2 Equipment

The HomeNet project aimed to make the Internet easily available und to reduce the most important barriers to usage, such as cost. A Macintosh Performa 475 computer with 12 MB of RAM and a 160 MB hard disk, a color monitor, a 14.4kbps modem, a printer and an additional phone line for Internet use only were given to each household. Each computer was preloaded with software such as word processor, spreadsheet software, home finance package, games, and a software suite for electronic mail, news groups, and World Wide Web browsing. Hard and software was given to the households for no charge.

At the time the HomeNet project was started, this was state-of-the-art equipment. By providing project participants with this equipment, and by not charging for it, the HomeNet project helped to overcome technological and economic barriers to use.

2.2.3 Variables Measured

The selection of variables depends heavily on the research question one wants to ask. Because this dissertation deals with issues that are relevant for electronic commerce and public policy, issues such as the digital divide and Web user profiling are addressed. In this regard, we used demographic data on age, race, sex, and family role, sociographics, and psychographics of individuals.

Gender

We wanted to test if there is a gender gap in adopting new technology. Therefore, gender was one of the used variables. Summary statistics of the sample reveal that 49% of the HomeNet users were male, 51% were female.

---

[3] The number of people in the project changed over time. In 1995, the project started with 48 families. As of March 1997, 100 families were part of the trial.

Race

We also wanted to test if there is a race gap in adopting Internet technology. In this regard, race or ethnical background was another variable used. According to summary statistics, 72% of the HomeNet users were white Caucasians, whereas 28% of the users belong to a minority group.

Age

A generation gap may exist in adopting new technology. Therefore, age was the next variable measured. 74% of the people in the HomeNet project were adult[4]. Tab. 1 shows some summary statistics on the age of the individuals.

Tab. 1: Age of individuals in the HomeNet sample

	Average	10th percentile	50th percentile	90th percentile
Age of individuals at time of joining the project	31.91	10.80	34.00	52.20

Role in the family

According to Kraut et al. [Kraut96b], teenagers – sons and daughters - turned out to be much heavier Web users than their parents. Thus, the family role of individuals was measured. Tab. 2 depicts the statistics with respect to family role of individuals.

[4] >= 21 years

Tab. 2: Family role of individuals in the HomeNet sample

Role in the family	Percentage
Mother	25.90%
Father	20.14%
Daughter	20.14%
Son	17.27%
Other	16.55%

Computer Skill Level

Another important variable used is the computer skill level of individuals. A lack of training or education may be accountable for low Web usage. Tab. 3 shows statistics on the self-reported computer skill level of individuals in the HomeNet sample.

Tab. 3: Computer skill level of individuals in the HomeNet sample

	Average	10^{th} percentile	50^{th} percentile	90^{th} percentile
Computer skill level[5]	3.43	2.20	3.60	4.68

Household income

Studies on the digital divide as is described in chapter 7 suggest that household income has an effect on whether or not individuals can afford a computer, which affects subsequent Internet usage. The HomeNet project aimed to overcome economic barriers to use. The finding from the HomeNet project that neither household income nor education predict Internet use, does strongly suggest that if economic barriers to Internet

[5] self reported 5-point scale ranging from 1 (low) to 5 (high)

access were removed, people across socioeconomic lines would use the Internet. In this regard, household income was measured in order to find out if household income does still affect Internet usage, even if different measures of Web usage are applied. Tab. 4 shows descriptive statistics on household income of individuals in the HomeNet sample.

Tab. 4: Household income of individuals in the HomeNet sample

	Average	10^{th} percentile	50^{th} percentile	90^{th} percentile
Household income	$54,410/year	$27,500/year	$52,500/year	$85,000/year

Also, Web usage, phone usage, and e-mail usage were measured. The next section deals with data collection issues.

2.3 Data Collection

The HomeNet data was assembled from five sources: [Kraut96a]

1. Pre-trial, bimonthly, and post-trial questionnaires
2. Computer-generated use records of electronic traffic, newsgroups read and posted to, Web sites visited, and time on the Internet
3. An archive of HomeNet newsgroup messages
4. A log of help requests
5. Home interviews

Clearly, these are rich and valuable sources of data. For this dissertation, we have used the first two data sources. We used data from questionnaires (demographic data measuring age, race, sex, and family role, i.e., mother, father, son, daughter, and others such as uncle and aunt, sociographic data, i.e., household income, and psychometric data, i.e., computer skill level) of all households. Also, we linked these data with computer generated usage data on Web usage, phone usage, and e-mail usage.

Collecting and extracting computer-generated use data is a non-trivial task. Therefore, it is crucial to understand the fundamentals of the technology behind the collection of such usage data. Specifically, Web log files that capture the Web clickstream of individuals deserve special attention.

HTTP, the Hypertext Transfer Protocol, is the underlying protocol of each access to a Web page. The users on the client side request a document at a specific Internet address (represented by a 'URL' – Uniform Resource Locator). The Client Software (Web browser) initiates the connection to the Web server on which the requested document is stored. The server sends the requested file back to the client. See also [Köhntopp00].

In contrast to the traditional method of getting information on user behavior, such as questionnaires, the clickstream is automatically generated. Such Web use clickstream can be collected at the client side[6], at a proxy server[7], or at the Web server[8].

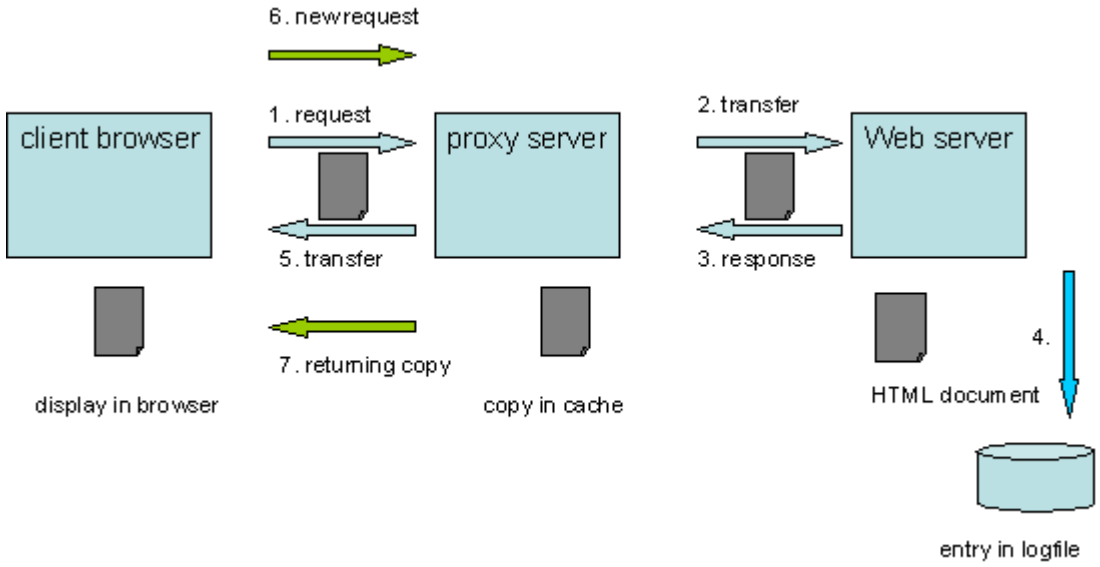


Figure 3: HTTP caching [Wilde99]

However, caching of Web pages might have an effect on the quality of Web logs. A cache may exist at the client side, proxy server, or Web server. When a document is requested for the very first time, the Web server adds an entry to the logfile and sends the document

---

[6] A client is an application that runs on a workstation (client part of the client-server architecture) and relies on a server to perform some operations.

[7] A proxy server is a server that sits between a client application, such as a Web browser, and a server. It intercepts all requests to the server to see if it can fulfill the requests itself. If not, it forwards the request to the server.

[8] A Web Server is an application or computer that delivers Web pages. It sends files/documents (e.g., HTML documents) to Web clients upon request. Each access is stored in a Web log file. Such a log file represents the entire protocol of Web server activity.

to the client. A copy of the document may be stored at the proxy server or even directly at the client's machine (see Figure 3).

The process of logging user actions at the Web server is as follows: Every Web server stores a variety of data on Web usage in Web log files. Each line in this log file represents an access to the files provided by this Web server. However, if the client requests the same document once again, the request may not reach the Web server because the wanted document is delivered by the client's cache or the proxy server. In this regard, multiple requests of the same document may not appear in the Web server's log file.

Client-side logging uses software installed on all the client machines that monitors user actions. Client-side logging provides higher accuracy in measuring Web user activities, because Web or Proxy server logs do not capture Web access from the client's local cache, which contains Web pages requested via the browser's Back and Forward buttons. However, collecting Web usage data on the client side requires substantial effort. An example of client-side logging is described in [Montgomery00]. See section 3.3.2 for a detailed discussion of this issue.

HomeNet used its own proxy server to capture user behavior. Logging user actions at the proxy server is similar to logging at the Web server side. However, it has some crucial advantages such as avoiding the problem of pages cached at the proxy server. Also, because all HomeNet participants had to use the same proxy server, all visits all users made to any Web server were recorded. This is in contrast to Web server logs that by nature only capture Web usage at one Web server. This also reveals the main drawback of logfiles that are collected at the Web server: As reported in [Adamic01] and confirmed in section 5.3, the distribution of users with respect to Web sites visited can be described mathematically as a power law distribution. Millions of users go to a few sites, paying little attention to millions of others. If one were to select a group of Web sites at random, and count the number of users visiting this site, the majority would be smaller than average. There is a discrepancy between average and typical behavior that is due to the skew in the distribution. Therefore, Web use logs should preferably be collected on the proxy server or the Web browsing client system.

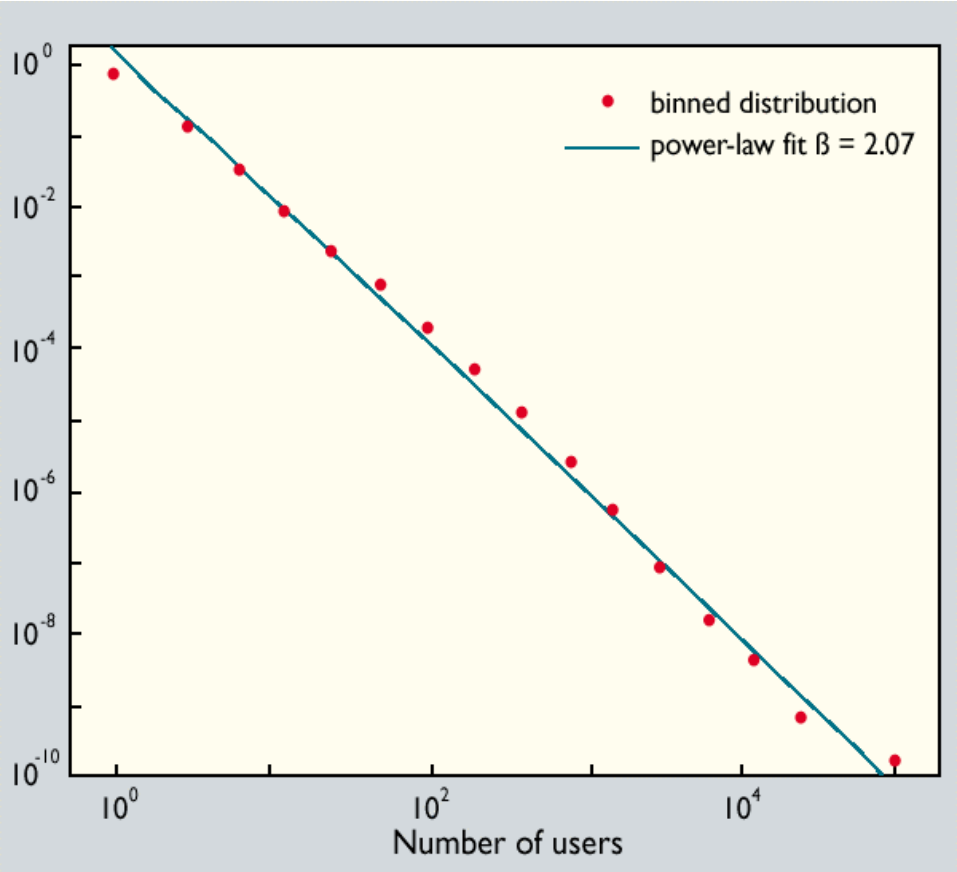


Figure 4: The power law distribution of Web sites [Adamic01]

The available fields of the HomeNet clickstream data set collected at the proxy server are as follows:

1. Unique user ID
2. Time and date of the user activity
3. URL accessed, which is comprised of
   a) Domain / Web site accessed (e.g., 'www.yahoo.com')
   b) Document on the Web server accessed (e.g., '/search/index.html').

Figure 5 depicts an excerpt of the HomeNet logfile. Every record in the clickstream represents a hit of a Web user.

```
197022;01.08.1996 15:15:00;GET;http;home.netscape.com;/escapes/search/search5.html
197022;01.08.1996 15:16:00;GET;http;www.netzone.com;/~dburns/zizza2.htm
197022;01.08.1996 15:20:01;GET;http;www.webcrawler.com;/
197022;01.08.1996 15:21:01;GET;http;www.mindspring.com;/%7Engan/9stargate.html
197022;01.08.1996 15:21:03;GET;http;www.compass-ent.com;/stargate/1991cal.html
197022;01.08.1996 15:22:05;GET;http;www.stargate.com;/
197022;01.08.1996 15:24:02;GET;http;www.stargate.com;/magazine-toc.html
197022;01.08.1996 15:24:03;GET;http;www.stargate.com;/sep96/toc.html
197022;01.08.1996 15:25:01;GET;http;www.stargate.com;/play/max-sep96.html
197022;01.08.1996 15:26:04;GET;http;www.stargate.com;/play/datasheet-sep96.html
197022;01.08.1996 15:57:03;GET;http;www.stargate.com;/play/mx/sep96/hold.html
161901;01.08.1996 15:58:02;GET;http;home.netscape.com;/home/whats-cool.html
```

Figure 5: HomeNet logfile example

2.4 Clickstream Data Extraction - Web Logfile Analysis

2.4.1 Fundamentals of Web Logfile Analysis

Organizations conductioning e-commerce can greatly benefit from the insights gained from the analysis of the clickstream of Web users [Kohavi01]. In general, an analysis of the Web user's clickstream may reveal information on Web usage that is relevant from a business perspective [Schwickert01]. Such analyzing of a Web usage clickstream from a business perspective involves issues such as electronic customer relationship management (eCRM)[9], because the clickstream describes the relationship of users and a Web site [Schaarschmidt01]. In this regard, a careful analysis of detailed data on Web user behavior in general and the clickstream in particular is a necessary precondition of eCRM.

However, the clickstream itself is of little use for business related purposes, such as marketing or customer relationship management. One has to ask the right questions and

[9] CRM deals with the relationship between customer and company and involves planning, coordination, and control of business activities. It is an integrated approach that aims to optimize all the customer related processes such as marketing, sales, and services across channels. Retaining existing customers is cheaper than winning new customers. Understanding and anticipating customer needs and aspirations is one of the crucial goals of CRM that eventually helps to get profitable customer relationships. Therefore, user segmentation is one of the key elements of CRM [Schaarschmidt01]. The term eCRM refers to the management of customer relationships using new technologies for better serving of the customer.

analyze the clickstream according to these questions to get helpful insights in user behavior (e.g., eCRM aims to classify users with respect to specific questions that are relevant from a business perspective [Schaarschmidt01]). Also, the sheer amount of data (i.e., 65 million page views at yahoo.com per day in 1997 [Yahoo99a]) requires a well-considered extraction of data and the right measures of Web usage.

When measuring Web usage in general and Web site success in particular, it is important to agree upon a set of commonly used measures, to make comparison between Web sites possible. The key terms of Web log file analysis are as follows:

- User
- Visit (or session)
- Page view
- Hit

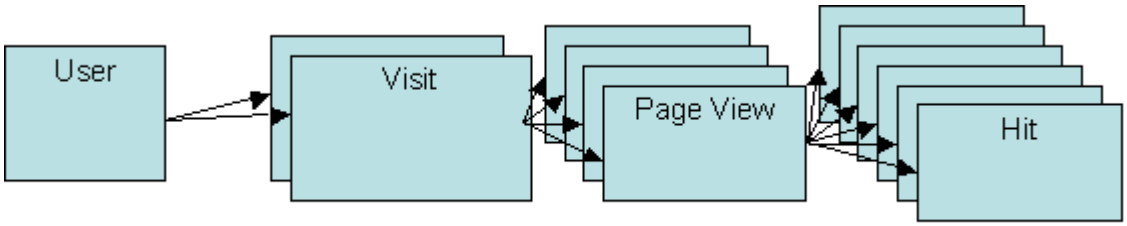


Figure 6: Understanding the difference between users, visits, page views, and hits [Cutler00]

User

A **user** is the individual accessing a Web site. In this regard, it is important to assign specific Web requests to users. Schaarschmidt et al. [Schaarschmidt01] distinguish three types of users:

- Anonymous user (only IP address given)
- Identified user (User is identified, for example by cookie id[10], name and other data are unknown)

[10] A persistent cookie is a unique identifier assigned by the Web server to each client accessing the site (or other sites associated with the same domain) for the first time. This identifier is stored on the client side and transmitted back to the server upon subsequent visits to the server by the same client [Berendt02a]. In this

- Known user (user is identified and name is known)

Known and identified users are particularly useful from an eCRM perspective. Note that the HomeNet data set provides the entire clickstream of known users. Moreover, it provides demographic and other kinds of data. In this regard, [Schaarschmidt01] points out that linking clickstream data with demographic data is crucial for eCRM. This dissertation aims at integrating clickstream and external data such as demographic data and analyzing it from a business perspective (which is the very purpose of eCRM in general) and from a public policy perspective. Therefore, the techniques and specific results of this dissertation provide a foundation for eCRM, whose typical tasks include identifying Web users and their characteristics, and identifying users that may discontinue using a service. Identifying users that may leave the site forever involves the issue of churn and loyalty as well as the issue of developmental trajectories discussed in subsequent chapters.

Visit / page view

A **visit** or session is the act of using the Web. However, identifying visits from Web logs is difficult and an issue for ongoing research (see also section 2.4.2: 'Pitfalls of Web Log Analysis'). A visit usually consists of one or more **page views** of documents.

Hit

Each page may consist of a variety of objects such as inline images, other HTML documents, or video streams. Every access of a file on a Web server is stored in the logfile. Therefore, a page view produces usually more than one **hit** on the Web server.

Not all the available measures of Web usage are interesting from a business or policy perspective. For example, success of a Web site is often measured in number of hits at this Web site, which equals number of lines in the logfile. Clearly, the size of the logfile is a very inaccurate measure of Web site success. Schmitt et al. [Schmitt99] point out that 'using hits […] to judge site success is like evaluating a musical performance by its volume'.

---

regard, by letting the Web server store small amounts of data at the client side, cookies provide a means for better identification of users. However, this comes at the price of less user privacy.

Businesses are rather interested in measures such as number of users and number of customers. Also, the click-through-rate[11] [Janetzko99] and the duration of users at specific pages are useful measures of Web usage. However, logfiles only provide a collection of pages accessed by clients. Getting information relevant from a business perspective by using logfiles is comparatively difficult. For example, the time stamps provided by the logfiles do not necessarily represent time actually spent on a page. Users might walk away from the computer and continue viewing the page later on. These and other pitfalls associated with Web log analysis are discussed in section 2.4.2. This dissertation aims to overcome these problems by developing and applying measures of Web usage that are relevant from a business and policy perspective. For example, the number of unique Web sites accessed as a measure of a user's interest in the Web is developed in chapter 3. User sessions and the duration of these sessions are identified in chapter 4. Measures of loyalty and Web portal utilization are developed in chapter 5 and chapter 6.

2.4.2 Pitfalls of Web Log Analysis

Kohavi [Kohavi01] reports some of the pitfalls associated with Web log analysis (based on the information provided in the HTTP header):

- Web logs do not identify sessions or users.

  HTTP is stateless[12]. Therefore, a 'session' does not exist at the level of the Web or proxy server. Identifying sessions from Web logs is an active research topic [Cooley99, Berendt01, Catledge95]. Techniques[13] rely on cookies, time, and client IPs.

---

[11] The click-through-rate is the proportion of users exposed to banner advertisements who actually followed the hyperlink provided by the banner.

[12] HTTP 1.0 is a stateless protocol. After each file transfer the connection is lost and rebuilt upon the next request. However, if requests come from the same client and if requests occur condensed in a specific period of time, it is reasonable to assume that they belong to the same visit. The time-out maybe between 5 und 30 minutes (see [Bensberg99]).

[13] There are a lot of techniques that help to identify unique users from logfiles (for example, an easy but inaccurate way is to use the client's IP address as a unique user ID. Simply using the IP of the requesting client for the purpose of identifying the user is an inaccurate way of user identification. Many users may use the same client machine. Also, many client machines may use the same proxy server, which substitutes the client machine's IP with its own IP address. In this regard, HTTP request may appear as requests from one machine. If this machine is a proxy server, this IP may be representing a variety of client IPs that use this proxy server (see [Garfinkel99]). Frames and 'Virtual users' such as robots make logfile analysis more

Problems may arise due to proxy or client caches, IP reassignment, and browsers rejecting cookies.

We deal with the issue of session identification by applying heuristics in chapter 4. User identification is not an issue because the HomeNet data sample provides unique user IDs and additional data of all the users in the sample.

- Web logs need to be conflated with transactional data.

  Metrics such as the product conversion rates [Gomory99] depend heavily on such additional transactional data. However, because this dissertation is considered basic research, this issue is beyond the scope of this work and subject to future research.

- Web logs lack critical events.

  Events such as ‘ending a session’, ‘leaving the computer’, and ‘abandoning a page or a shopping cart’ are not easily computable from Web logs. Specifically, identifying visits is a non-trivial problem due to the very nature of the Internet architecture. As noted above, we approach the problem by applying heuristics in chapter 4 on pp. 62 ff.

- Web logs contain URLs, not the semantic information of what the URLs contain.

  There is no standard approach to deal with this fundamental problem. One possible approach is to manually add semantics to URLs, e.g., by manually assigning Web sites to categories (see [Berendt02b] for other approaches), such as conducted in chapter 3: ‘Saturation of Lay Web Usage’.

- Web logs lack information for modern sites that generate dynamic content.

  Dynamically generated Web pages make it harder to extract information from Web logs. Often, URL templates are reused to present different information. We deal with this issue by excluding dynamically created pages from the analysis. At the time the HomeNet data was collected, only a small percentage of Web servers generated dynamic content.

- Web logs contain redundant information.

  The sheer amount of data in Web logs makes an analysis more difficult and requires substantial computational effort. Therefore, we consider only data in Web logs that is relevant from a business or policy perspective, such as pages viewed by users. See also section 2.5: 'Normalizations and Assumptions’ on pp. 35 ff.

---

complex. Puscher [Puscher00] discusses ways of avoiding skewed results of logfile analysis that result from requests by spiders, crawlers, or bots.

- Web logs lack important information that can be collected using other means.

  The HomeNet data sample provides lots of additional data that was collected to overcome this pitfall. For example, demographic data was collected. However, some types of data are very difficult to collect, e.g., data on user intentions.

Other pitfalls associated with Web log analysis include the problem of HTML frames. Frames often contain content that the user did not explicitly request, such as banner advertisements. Nevertheless, these requests are recorded in the logfile. However, as in the case of dynamically generated content, the use of frames became popular after the data for this analysis was collected.

In summary, there are many problems associated with Web log analysis that this dissertation aims to overcome. The discussion above speaks for the use of proxy-side logfiles combined with additional data on user characteristics, such as provided by the HomeNet project.

2.4.3 Software Tools Used

There are lots of commercial software tools for Web log analysis available. Leading software companies on the Logfile-analyser market are Webtrends® [Webtrends] and Exody® [Exody]. More complex software such as Accrue insight® [Accrue] offers the linkage of usage data and demographic data. However, because these software programs do not offer the flexibility needed for this analysis, we did not use these software tools. Instead of using commercial tools, we imported the clickstream into an oracle® 8.1.6i database [Oracle], which was queried by self-coded Java and Visual Basic® software that used ADO[14] [ADO] and JDBC[15] [JDBC] interfaces. Also, for parts of the analysis in chapter 4: 'Analyzing Web Sessions' on pp. 62 ff., the free software 'analog' from the SourceForge-Projekt [SourceForge] was used. In order to use this software, the

[14] Microsoft® ActiveX® Data Objects (ADO) is Microsoft's interface to multiple types of data. It provides consistent access to data for 1-tier to n-tier client/server and Web-based data-driven solution development. It enables client applications to access and manipulate data from a database server through an OLE DB provider. OLE DB is Microsoft's data access paradigm, which provides access to any data source, including relational and non-relational databases. ADO supports key features for building client/server and Web-based applications.

[15] JDBC™ is an API that provides cross-DBMS connectivity to a wide range of SQL databases.

clickstream had to be converted to the Common Logfile Standard (CLF) [CLF]. This was done by running self-coded Perl scripts [Perl].

The data collection and extraction methods described above were used to create Tab. 5, which reveals first summary statistics on Web usage (and phone usage and e-mail usage) of individuals in the HomeNet sample.

Tab. 5: Summary statistics on Web usage, e-mail usage, and phone usage of individuals in the HomeNet sample

	Average	10th percentile	50th percentile	90th percentile
Number of distinct sites (domains) visited monthly	6.39	2.15	5.50	11.43
Number of pages viewed monthly	18.27	5.00	15.00	34.23
Phone usage[16]	4.14	3.00	4.00	5.00
Connection hours weekly	2.15	0.03	0.78	7
Mails sent[17]	1.62	0.00	0.39	4.91

As reported in [Kraut96a], Internet usage varies widely. For example, the average number of distinct Web sites visited per month is 5.5, whereas the 10th percentile is only 2.15 and the 90th percentile is 11.43 distinct Web sites visited. Established technologies such as the telephone show much smaller degrees of variance. Analyzing the usage patterns of individuals more deeply than what is shown in Tab. 5 is the purpose of this study.

2.5 Normalizations and Assumptions

The data set on which this study is based on consisted of 139 users performing 1,187,325 http requests between 11-6-1995 and 4-28-1997. In conducting the analysis, the following

[16] self reported 5-point scale ranging from 1 (low) to 5 (high)

[17] self reported 5-point scale ranging from 1 (low) to 5 (high)

normalizations and assumptions are made. First, the data set included many http requests that users did not explicitly perform, such as requests for image files, which are automatically generated by the Web browser ('hits'). Also, http invocations using the common gateway interface to external programs are often not explicitly requested by the user but instead are automatically loaded (dynamically generated content)[18]. Including these download requests would incorrectly inflate Web usage. For example, 66.7% of the downloads were inline images, 5.13% of the requests were downloads using the common gateway interface (cgi), and 10.0% were downloads of other types, such as music or video files. Therefore, an important key measure of this dissertation is the number of pages viewed by individuals. As noted in the previous section, 'page views' are requests for documents (rather than requests for inline image, movie, audio file, etc.). In this regard, records that did not have one of the following suffixes: .htm, html, .jsp, and .asp, were removed from the database. Records that point to a directory rather than to a file remained in the clickstream database, because Web servers automatically respond to these request by sending a standard document, such as 'index.html'. Removing all hits that were not page views reduced the size of the data set by 81.9% to 214,818 downloads.

Next, *domains,* which are identical except for the prefix, (e.g., 'www.yahoo.com' and 'yahoo.com') were treated as the same domain. Also, domains, which have the suffix that indicates an explicitly requested port, such as 'yahoo.com:80' were truncated to 'yahoo.com'. Further, it is assumed that each domain represents a single Web site.

Finally, because users began their use of the Web at different starting dates and because of different individual durations of Web use in the project, there was the issue of sparse or missing data at the end of each individual's monitored period of Web usage. Therefore, we analyzed the evolution of Web usage over an 8-month period of time beginning with each individual's starting date. Cutting off sparse data at the end of the period of observation reduces the size of the data set by 6.9%. After these normalizations, the data set consisted of 133,421 page views.

---

[18] For instance, an HTML page containing form elements such as those used to submit credit card numbers have references to cgi-bin programs. Visits to the HTML page are counted but calls to the cgi-bin programs that are invoked automatically to process the entries submitted using the form are left out.

2.6 Quality and Representativeness of the Sample

The quality of this residential data sample is highlighted by the high percentage of people accessing the Internet at home. Points of Internet access include:

- Access at home
- Access at work
- Access at schools
- Access at public libraries and community centers
- Mobile access[19]

Already in 1999, 22.2% of all Americans used the Internet at home, and 17% used it at some site outside the home [NTIA99b]. A substantial share of overall Web usage is residential Web usage. This underlines the importance of measuring Internet usage in residential settings.

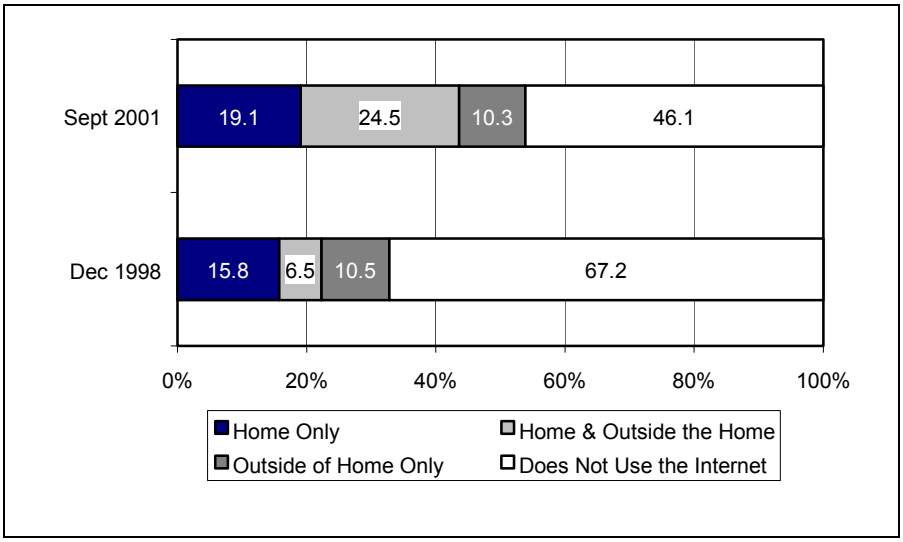


Figure 7: Internet use by location as a percent of United States population, 1998 and 2001 [NTIA02]

[19] As reported in Ferguson et al. [Ferguson01], mobile access gains importance very slowly. Falling stock prices of telecommunication companies and depreciated cost of UMTS licenses are an indicator of slower than expected adoption of mobile technology.

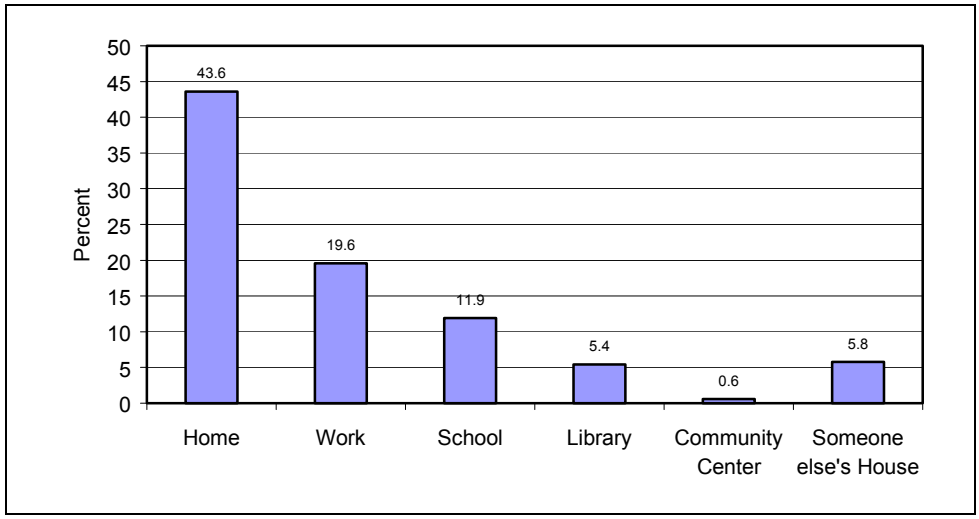


Figure 8: Internet use by specific location as a percent of United States population [NTIA02]

With the exception of the recent work by Montgomery and Faloutsos [Montgomery00], other studies (e.g., [Tauscher97a], [McKenzie01a], and [Catledge95]) rely on highly non-representative samples (e.g., individuals who worked or studied in computer science departments). This study relies on a sample of households that is more closely representative of the general population.

The period of observation and the tenure of the individuals in the panel, eight months, is far longer than in prior studies. An extended period of observation is required to calibrate credible patterns of change in Web use. Eight months may still be too short a period of time to measure fully the development of Web usage behavior but it is clearly better than the much shorter study periods of prior studies.

Event though the data is comparatively dated, it is still considered a valid data source from which implications for the new Millennium can be drawn. Access technology and the Web itself did not dramatically change. Technological breakthroughs such as ubiquitous computing are still far from being deployed. Even in 2001, years after the start of the HomeNet project, [NTIA02] reports that personal computers with modem capability are still the mode of choice for Internet access (see Figure 9). Therefore, it is strongly believed that the results from 1995-1997 are still valid for today since the access technology has not changed much. However, it is reasonable to assume that this will change eventually. Mobile devices will ensure ubiquitous access and flat rates will change Web usage patterns (See also future work issues discussed in chapter 8 on pp. 133 ff.).

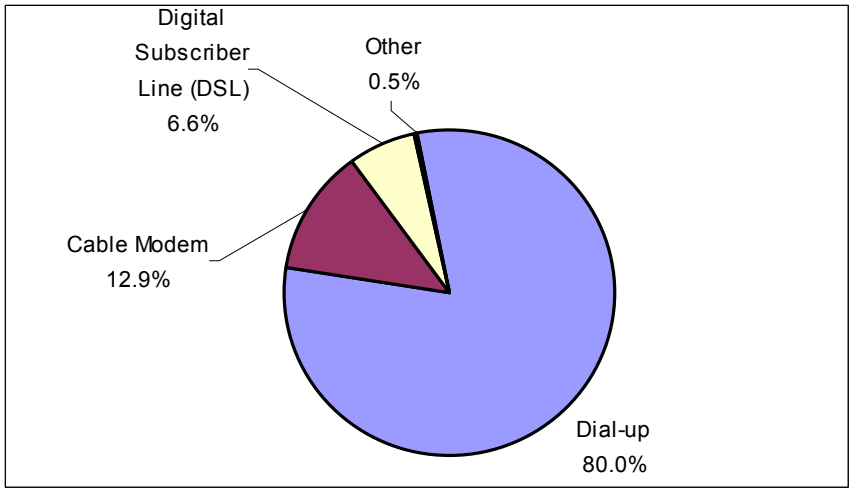


Figure 9: Home Internet connection type, 2001 as a percent of Individuals using the Internet at home [NTIA02]

Using data that was collected between 1995 and 1997 also offers some advantages compared to using more recent data. In 1995 it was easier to find 'new users' who had not used the Internet before than it is today. Also, 1995-1997 was exactly the period of time in which the Web became popular, the number of Web sites offered grew exponentially, and many search engine transitioned to Web portals. Because measuring the effect of an exponentially growing number of Web sites offered and Web portals on Web usage patterns of new Internet users is the very purpose of the analysis in subsequent chapters, this data set provides the necessary quality needed for these studies. Also, using data from 1995-1997 avoids the problem of Web log analysis with frames and dynamically generated content, because they were rarely used at that time. Moreover, because e-privacy became an issue after the HomeNet data used for this dissertation was collected[20], there was less danger of 'Hawthorne Effect' modifications to subject behavior [Mayo33]. This is usually a problem if subjects are aware that their actions are being logged. Before 1997, the awareness of being logged among individuals was comparatively low.

[20] At the time the data used for this dissertation was collected, few people cared about privacy. However, already in 1998, studies conducted by the American Management Association show that 40% of the American corporations engaged in some kind of intrusive employee monitoring [Doyle99]. Such monitoring includes checking of e-mail and Web usage. Therefore there is a tendency towards a loss of privacy.

3 Saturation of Lay Web Usage

3.1 Introduction and Motivation

With the commercialization of the Internet, the Web has become a marketplace. Visits to given Web sites are considered an important measure of market share and success, and indeed, many Web sites have enjoyed a steady increase in the number of visits. Yet at the level of the individual user, little is known about the trajectory of change over time in the number of visits to Web sites.

Figure 1 on page 15 shows that over the period 1995 to 2000 there was an explosive growth in the number of Web sites available to users. The work presented in this chapter examines how users responded to the exponential growth in Web site availability. Specifically, we explore whether the increase in Web site visiting opportunities spurred an increase in the utilization rates of individual users. While there is much evidence of large increases in the number of users utilizing the Web [NUA], aggregate utilization reflects a combination of two distinct usage components - number of users and the intensity of their use. The objective of this chapter is to better understand the unfolding utilization rates of individual users. Specifically, we report results of an analysis of eight months of longitudinal data on individual level Web usage that was extracted from data collected as part of the HomeNet Project [Kraut96a]. A semi-parametric, group-based statistical method [Nagin99a] designed to identify distinctive trajectories of individual Web usage is applied to these data. This statistical methodology allows patterns of change in Web usage over time for individual users to be addressed directly. We focus on the analysis of the number of distinctive Web sites accessed per month as a measure of the user's interest in the World Wide Web. We thereby not only identify groups with different levels of usage, but also identify distinctive trajectories of the development of Web usage over time and provide demographic profiles of the identified user groups. The resulting trajectories are compared to the overall trend in the number of Web sites (see Figure 1 on page 15), which multiplied exponentially during the period of observation.

The chapter is organized as follows. Section 3.2 discusses the measurements of Web use and the statistical method applied to the data source described in chapter 2 (see page 19 ff.). Section 3.3 presents the results of this analysis. Section 3.4 discusses the implications of the results for Internet marketing strategy and public policy as it pertains to the digital divide and deals with limitations of this study and future work issues.

3.2 Measurement of Web Use and Statistical Method Used

3.2.1 Measurement of Web Use

There are several conceptually reasonable alternatives to measure Web usage. Broadly, they may be classified into frequency-based measures and time-based measures. Time-based measures use the time spent by an individual at a given Web site as an indicator of the utility received from utilizing the content of the site. In this chapter, we do not pursue time-based measures because constructing a measure that accurately reflected actual time spent actively interacting with a Web site is a non-trivial task. Users may download a Web site but only actively attend to the Web site for a small fraction of the duration over which the site was displayed (i.e., leave the computer unattended for hours or even days). Thus, for now we focus on frequency-based measures. Time-based measures will be discussed in detail in chapter 4: 'Analyzing Web Sessions' on pp. 62 ff.

Two frequency-based measures were analyzed - a count of the number of distinct Web sites visited by a given user per time period and total Web site visits per time period. The former measure does not count repeat visits to the same Web site whereas the latter measure does count such visits. Tab. 6 illustrates the calculation of these alternative measures of Web usage for a hypothetical user over a three-month period. In period 1, the user accesses a total of three sites. However, only two are distinctive because yahoo.com is visited twice. By this same counting logic, a total of four sites are visited in period 2 but only three are distinct.

Tab. 6: URL sets and number of distinctive Web sites accessed by a fictitious user

Month	URLs accessed	#distinctive Web sites
1	yahoo.com yahoo.com amazon.com	2
2	yahoo.com amazon.com excite.com amazon.com	3
3	yahoo.com yahoo.com	1

We use the count of distinct Web sites visited as the primary indicator of Web usage because of our interest in comparing the development of individual Web usage with the aggregate growth in Web site visiting opportunities. At the level of the individual the diversity of Web sites visited provides an indication of individual-level willingness to search the exponentially expanding set of visiting opportunities (see Figure 1 on page 15). However, because this measure does not count repeat visits and accesses to a given Web site, it is also important to examine total Web site accesses (e.g., page views) as an alternative utilization intensity measure. This permits an analysis of whether users that visit a few distinct Web sites in a time period are more intensive users of these sites than are users who visit a larger number of sites but use each of them less intensively. Further, the number of pages downloaded per site is an important measure with relevance for advertising online using banner advertisements. These advertisements are served as part of a downloaded Web page and priced per thousand impressions of the advertisement (also known as cost per thousand impressions or CPM, see [IAB] for more on online advertising).

Notice that the measures of distinct and total Web site accesses per month presented are not entirely accurate in the sense that they do not take into account Web usage behavior within distinct Web sessions. Therefore, chapter 4: ‘Analyzing Web Sessions’ on pp. 62 ff. further advances the measures of Web usage by following a session-based approach.

3.2.2 A Semi-parametric, Group-Based Approach for Analyzing Developmental Trajectories

Tab. 7 reports summary statistics on Web usage for 139 HomeNet users. The mean number of distinct Web sites visited per month is 32.66. Users commonly make repeated accesses to their favored Web site because the average number of page views per month is about 155 or 4.75 per distinct Web site visited. There is, however, much variation across users in utilization rates. The median number of distinct Web sites visited is only 10 sites per month, less than half the average. This implies a pronounced rightward skew in the population utilization rates, which is indeed reflected in the 90th percentile of distinct Web sites visited, 82.

Tab. 7: Summary information on Web usage

Overall number of page views	133,421
Average number of distinct sites visited / months	32.66
Average page views per month	155.15
Ratio of page views / site	4.75
10th percentile of distinct sites visits	0
Median of distinct sites visits	10
90th percentile of distinct sites visits	82
Users	139

Further, there may also be large differences across individuals in the unfolding of their utilization rates over time. This brings us to the central goal of this analysis—identification of the developmental course of Web usage across distinctive subpopulations. To this end we apply a semi-parametric, group-based methodology [Nagin99a] that was designed to identify distinctive trajectories of human development. In developmental psychology, a trajectory defines the developmental course of a behavior over age or time. Such trajectories might include groups of “increasers,” “decreasers,” and “no changers.”

Using finite mixtures of suitably defined probability distributions, the group-based approach for modeling developmental trajectories is intended to provide a flexible and easily applied method for identifying distinctive clusters of individual trajectories within the population and for profiling the characteristics of individuals within the clusters. Thus, whereas the hierarchical and latent curve methodologies model population variability in growth with multivariate continuous distribution functions, the group-based approach utilizes a multinomial modeling strategy. Technically, the group-based trajectory model is an example of a finite mixture model. Its parameters are estimated by maximum likelihood.

The fundamental concept of interest is the distribution of behavioral outcomes conditional on month of usage; that is, the distribution of behavioral trajectories denoted by $P(Y_i|month_i)$, where the random vector Y_i represents individual *i*'s longitudinal sequence of behavioral outcomes (i.e. Web usage) and the vector $month_i$ represents *i*'s month of Web usage when each of those measurements is recorded. The model assumes that this distribution arises from a finite mixture of unknown order *K*. The likelihood for each individual *i*, conditional on the number of groups *K*, may be written as:

$$P(Y_i \mid Month_i) = \sum_{j=1}^{K} \pi_j \cdot P(Y_i \mid Month_i, j; \beta_j),$$

where π_j is the probability of membership in group *j*, and the conditional distribution of Y_i given membership in *j* is indexed by the unknown parameter vector β_j. In most previous applications, β_j is a vector of regression parameters determining the shape of the group-specific trajectory.

For given *j*, conditional independence is assumed for the sequential realizations of the elements of Y_i, y_{it}, over the *T* periods of measurement. Thus, one may write

$$P(Y_i \mid Month_i, j; \beta_j) = \prod_{t=i}^{T} p(y_{it} \mid month_{it}, j; \beta_j),$$

where *p(.)* is the distribution of y_{it} conditional on membership in group *j* and the month of Web usage of user *i* at time t.

One valuable feature of the model is that it is easily adapted to accommodate different forms of data by an appropriate distributional representation of $p(y_{it}|month_{it},j; \beta_j)$. In this analysis the data is in the form of a count, whereby y_{it} measures the number of distinct Web sites visited by individual *i* in period *t*. As is conventional in the analysis of count

data, we assume that y_{it} follows the Poisson distribution. For the Poisson-based model it is assumed that, for each group j:

$$\log(\lambda_{it}^{j}) = \beta_0^{j} + \beta_1^{j} month_{it} + \beta_2^{j} month_{it}^{2} + \beta_3^{j} month_{it}^{3} \quad (1),$$

where λ_{it}^{j} is the expected number of occurrences of the event of interest (e.g. visits to distinct Web sites) of subject *i* at time *t* given membership in group *j*.[21] The model's coefficients - β^{j}_{0}, β^{j}_{1}, β^{j}_{2}, and β^{j}_{3} - determine the shape of the trajectory and are subscripted by *j* to denote that the coefficients are not constrained to be the same across the *K* groups. See Nagin et al. [Nagin99b] for further details.

A key issue in the application of a group-based model is determining how many groups define the best fitting model. One possible choice for testing the optimality of a specified number of groups is the likelihood ratio test. However, the null hypothesis (e.g. three components vs. more than three components) is on the boundary of the parameter space and hence the classical asymptotic results that underlie the likelihood ratio test do not hold [Gosh85].

Given these problems with the use of the likelihood-ratio test for model selection, we have followed the lead of [DUnger98] and use the Bayesian Information Criterion (BIC) as a basis for selecting the optimal model. For a given model, the Bayesian Information Criterion is calculated as *log(L) - 0.5*log(n)*(d)*, where *L* is the value of the model's maximized likelihood, *n* is the sample size, and *d* is the number of parameters in the model. [Kass95] argue that the Bayesian Information Criterion can be used for comparison of both nested and non-nested models under fairly general circumstances. When prior information on the correct model is limited, they recommend selection of the model with the maximum Bayesian Information Criterion. In even more recent work, [Keribin97] demonstrates that the Bayesian Information Criterion identifies the optimal number of groups in finite mixture models, a result specifically relevant for the mixture models demonstrated here.

---

[21] A log-linear relationship between λ_{it}^{j} and *month* is assumed to ensure that the requirement that $\lambda_{it}^{j} > 0$ is fulfilled in model estimation. Note also that the group-based specification accommodates population variation in λ. Such variation is the motivation for two important generalizations of the Poisson distribution, the negative binomial distribution and the zero-inflated Poisson distribution.

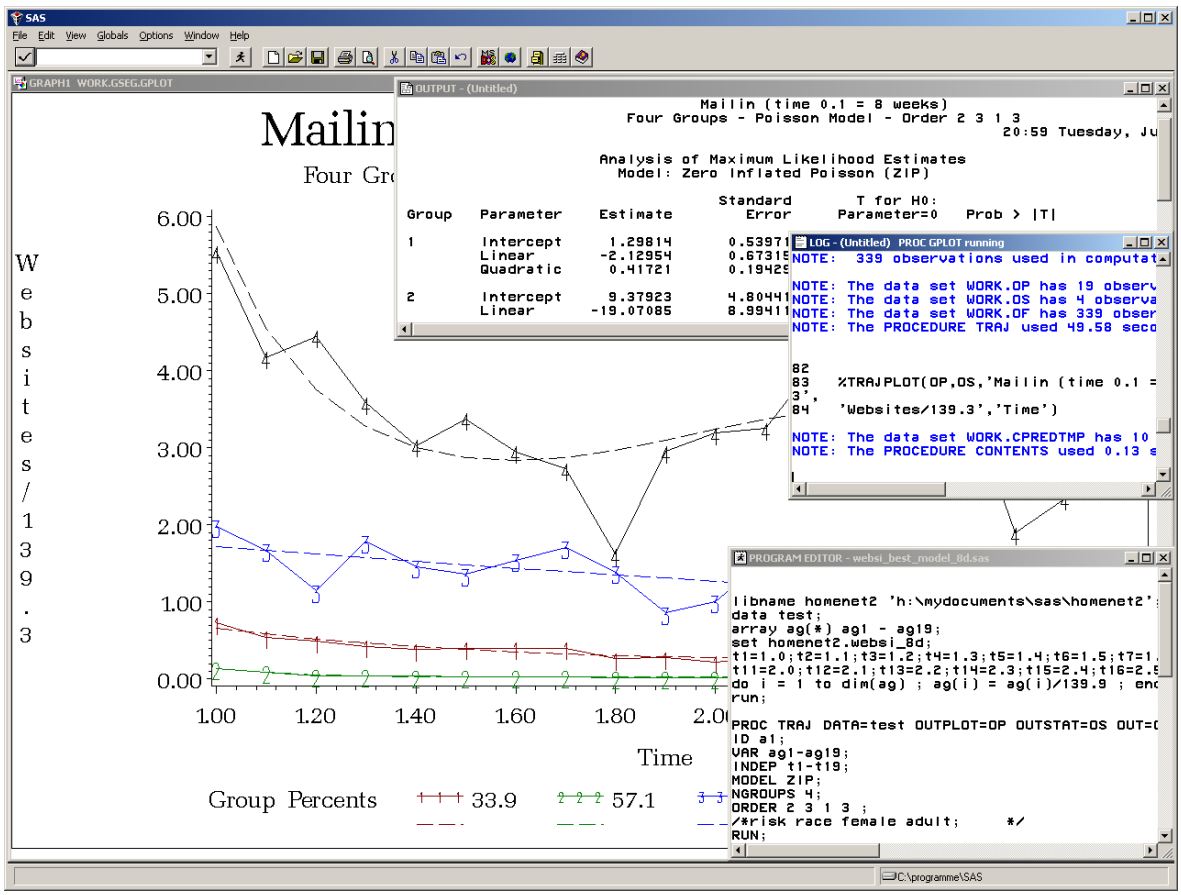


Figure 10: Graphical user interface of the SAS PROC 'TRAJ'

The software used to estimate these models is 'TRAJ', a customized SAS® procedure that was developed with the SAS® product SAS/TOOLKIT®. Figure 10 shows a screenshot of this software. Note that this software assists the user in all the necessary steps of modeling the data: It helps finding the model that fits the data best by calculating BIC scores, it estimates the parameters for a given model, and it computes the probabilities of group membership for each given individual as well as the demographic factors that discriminate the various groups. Furthermore, it uses the SAS/GRAPH® software for visualization of the trajectories. The procedure is a program written in the C programming language and is designed to interface with the SAS® system to perform model fitting. It accommodates missing data. Thus, individuals with incomplete assessment histories do not have to be dropped from the analysis. Also, time between assessment periods does not have to be identical across participants. The Software is available at [TRAJ] and is described in detail in Jones et al. [Jones02]. Alternative methods of model estimation are discussed in Nagin et al. [Nagin99a].

In summary, the method described above features the following characteristics:[22]

- Maximum Likelihood procedure based on mixture modeling
- Identifies distinctive trajectories of development
- Provides formal basis for determining number of groups
- Provides explicit metric for evaluating the precision of an individual's assignment to a group
- Provides group percentages
- Uses the data itself for model estimation

3.3 Results

3.3.1 Trajectories of Usage

Application of the group-based trajectory methodology to data from the HomeNet project revealed that the best fitting model, based on the Bayesian Information Criterion, clustered users into *four* groups. The existence of four groups confirms that Web usage over time is by far not distributed equally across users. Figure 11 depicts the actual and predicted trajectories of the four groups, which we label "*very heavy users*", "*heavy users*", "*moderate users*", and "*light users*". Because the utilization rates of the "very heavy user" group is so much higher than the other three groups, Figure 12 excludes the "very heavy user" group and depicts the predicted[23] and actual behavior of the other three groups. Tab. 8 shows the group percentages.

---

[22] However, the method does not comfort the user with brute-force capabilities that facilitate identifying the best model within a large range of possible models. In order to find the model with the largest BIC, as recommended by [Kass95] and [Keribin97], the user has to enter the SAS code for each possible model manually. For example, a 4-group model in which each group follows a third order polynomial requires 4^4=256 calculations. Therefore, we used a self-created source code generator that automatically generates the sources code for the SAS proc 'TRAJ'. The program is available upon request from christ@wiwi.hu-berlin.de.

[23] Predicted behavior is calculated as the expected value of each group's behavior and is computed based on model coefficient estimates. For this poisson-based model, this expectation equals the antilog of equation 1 on page 45. Actual behavior is computed as the mean behavior of all persons assigned to the various groups identified in estimation. As described in this section, the assignments are based on the posterior probability of group membership.

Tab. 8: Group percentages

light users	52.5%
moderate users	30.2 %
heavy users	13.7 %
very heavy users	3.6 %

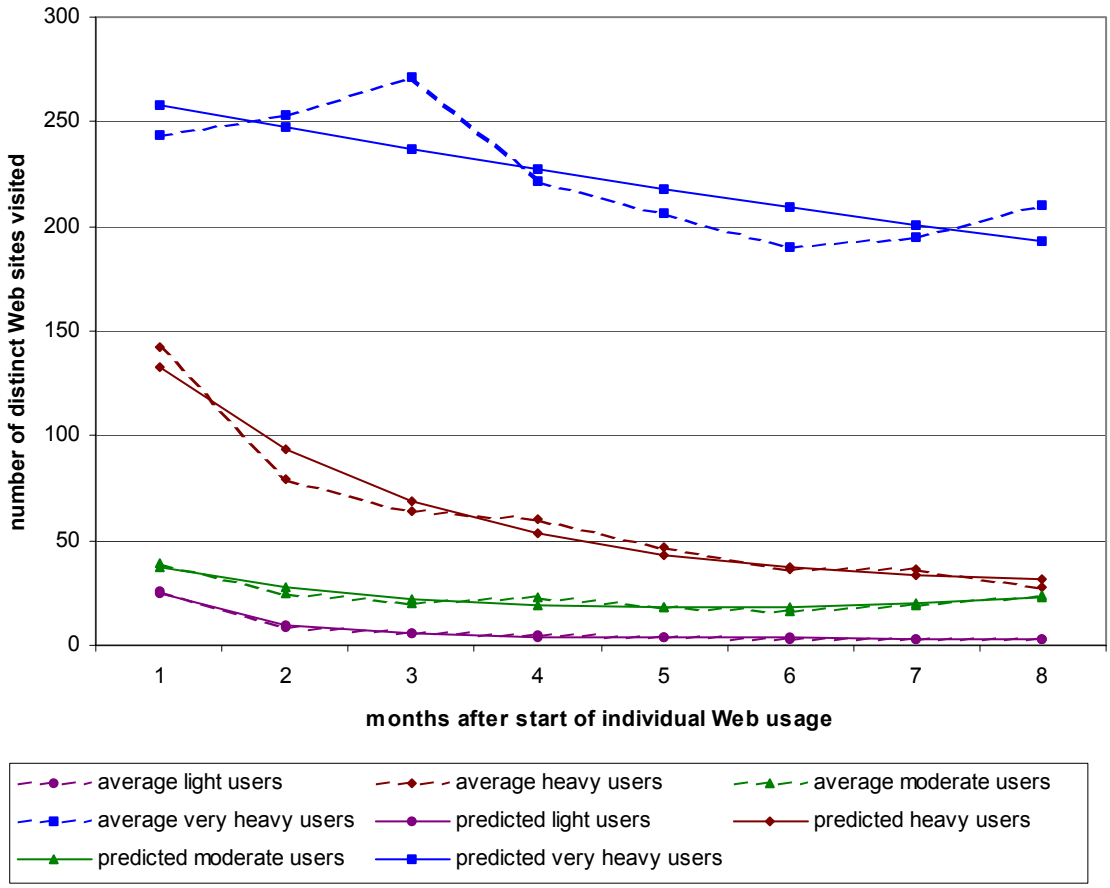


Figure 11: Residential use of the Web measured in number of distinctive Web sites accessed over time

Note that in contrast to the exponential growth in Web sites available as shown in Figure 1 on page 15, there is actually a decline in residential Web usage intensity as measured by number of distinctive Web sites accessed per month. But for a few initial visits to Web sites, the group of 'light users' is composed of individuals who make little use of the Web. This group is estimated to account for an estimated 52.2% of the sampled population. The

saturation level of 'light users' is only about 3 sites/month, indicating that this group did not find the Web particularly useful, following a short period of Web exploration.

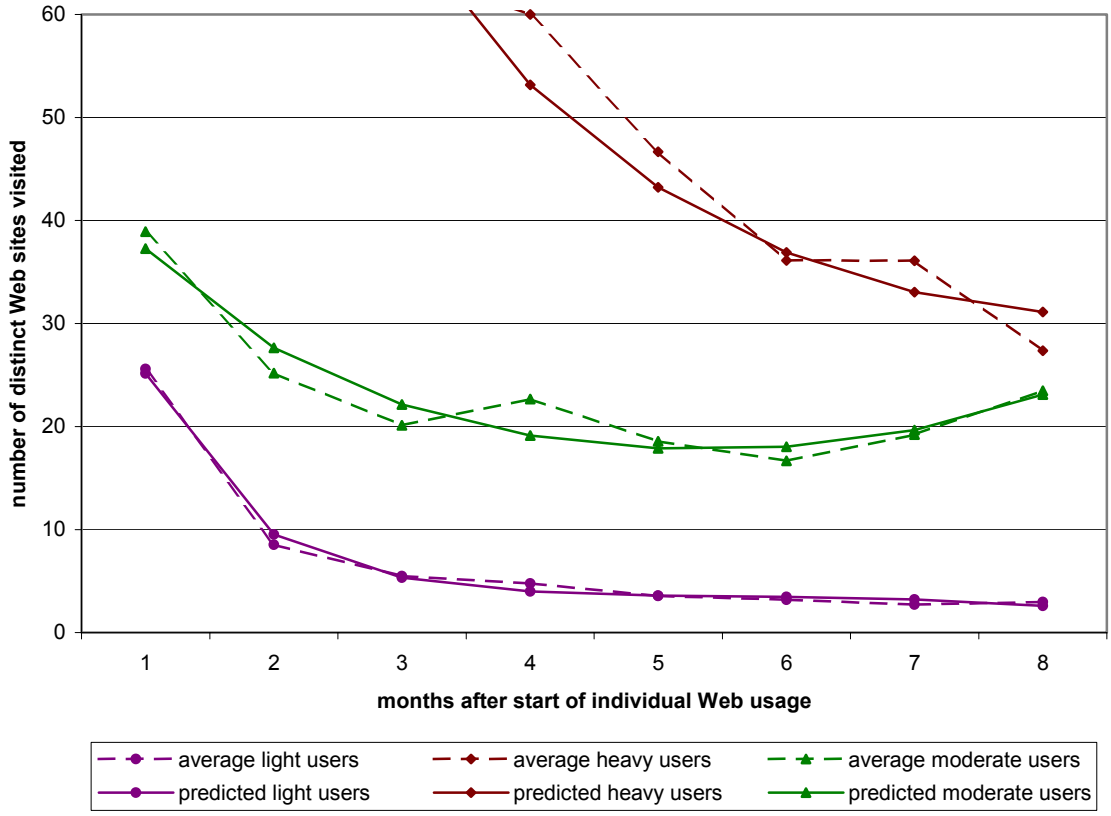


Figure 12: Number of distinctive Web sites visited over time; light users, moderate users, and heavy users only

The second group of individuals – moderate users – start Web usage at a higher level and follow a stable path in Web usage to a point of about 20 distinctive Web sites per month. This group is estimated to constitute 30.2% of the population.

The third group – heavy users who account for about 14% of the overall population – initiates Web usage at a high level of 140 distinctive Web sites per month. However, thereafter their utilization declines quickly to a saturation point of 33 distinctive Web sites per month, which is very close to the saturation point of the group of moderate users. Thus, while moderate users and heavy users differ in their initial Web usage, they converge to about the same utilization level in the long term. Finally, a 'very heavy user' group was identified and is estimated to make up 3.6% of the overall population in the

HomeNet sample. This group consists of users who started at a very high level, over 250 sites per month and who settle into a usage rate of about 200 sites per month.

In summary, all the groups appear to reach saturation in their extent of Web usage as measured by the average number of distinctive Web sites visited per month. For 52.5% of the population called ‘light users’, the saturation level is at a nominal level of usage of about 2 to 3 sites per month. For ‘moderate users’ the saturation level is about 20 distinctive Web sites per month. After initial heavy utilization of the Web, ‘heavy users’ tend to visit about 33 distinctive Web sites per month. A small minority of ‘very heavy’ users has a saturation level that is about 200 distinctive Web sites per month.

We consider these trajectories ‘learning curves’ of Web usage. The purpose of this analysis was to test whether these learning curves tracked the rapid increase in number of Web sites available, and the commercialization of the Net that occurred during the period 1995-1997. They did not. While the sampled households initiated their Web usage in this period of dramatic change in the Internet, no group followed a trajectory of increasing usage. On the contrary, all the groups follow a downward path, indicating that, after a period of ‘surfing around’ and ‘exploring’ the Web, residential users seem to limit their Web usage. The increase in available Web sites and the commercialization of the Web with its intended effort to appeal to users did not lead to an increase of Web usage at the individual level.

3.3.2 Intensity of Web Utilization

As shown in Tab. 7, the intensity of utilization as measured by number of page views is considerably larger than when measured in number of distinct Web sites. This indicates that individuals are making multiple visits to Web sites, which is desirable from the perspective of a Web site operator. Figure 13 depicts the distribution of page views by a trajectory group over time. As with visits to distinct Web sites, the number of page views is stable or slightly declining over time, depending on group membership.

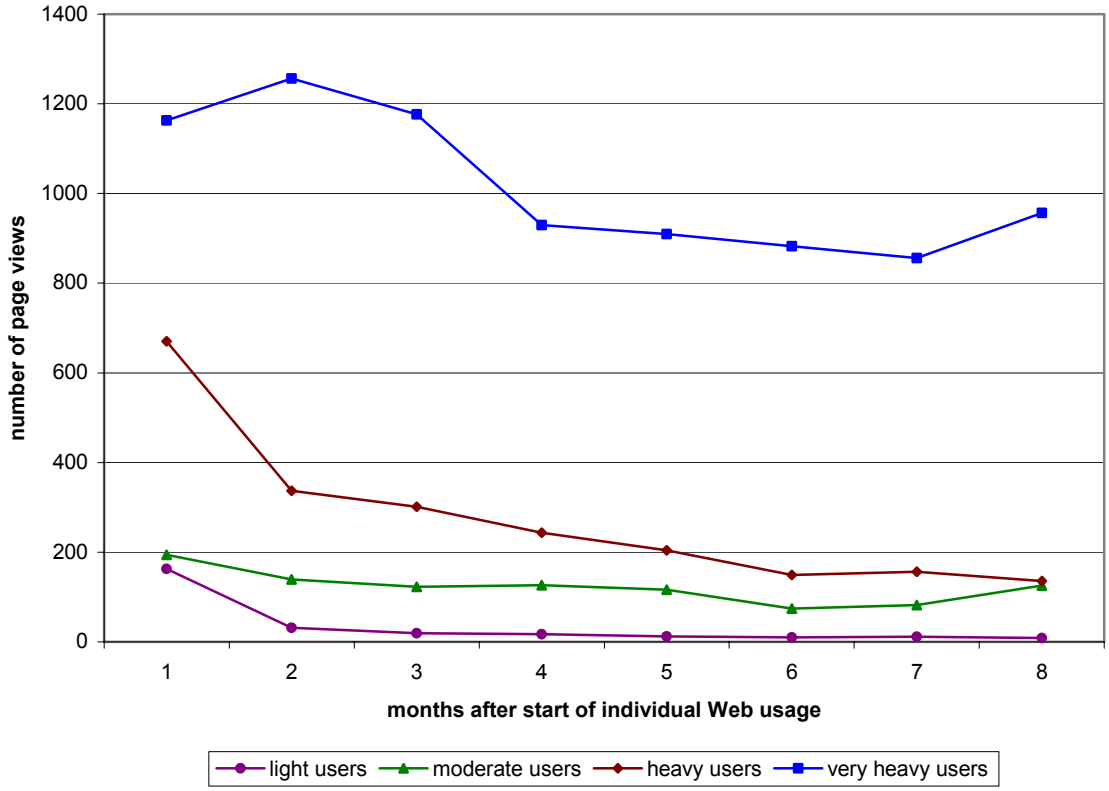


Figure 13: Distribution of monthly page views over time by trajectory group

The results confirm that there is no increase in Web pages viewed by individuals. However, this key result that all groups of users achieve saturation in their Web utilization (also measured as the number of Web viewings per user) is contrary to the key results of Montgomery and Faloutsos [Montgomery00] who found that Web usage as measured by the number of Web viewings per user is increasing for all types of users.

Before we move on, it is important to explain this difference. The tenure of individuals in the panel is an important distinction between this study and the results reported by Montgomery and Faloutsos. As noted in [Montgomery02], the number of months an individual spent on their panel has a pronounced leftward skew. 38% of the users spent between 1-3 months on the panel and 22% of the users spent between 4-6 months on the panel. Thus 60% of the users spent less than 6 months on the panel. This implies that they have high churn in their panel. Data on Web usage includes both, data from novice and experienced users. In contrast the HomeNet data set followed the browsing behavior

of *each user* over the entire 8 month period over which data was collected and pertains to steady state browsing behavior of users.[24]

In interpreting page views, it is important to keep in mind that individuals do not necessarily visit the same distinct Web sites from month to month. Indeed, there might be considerable churn in the specific Web sites visited from month to month. In this case, there would be limited overlap over time in the identities in the specific Web sites visited. Still while not a perfect measure of loyalty, the number of page views per site is an indicator of loyalty. Loyalty of Web users to Web sites is an issue of utmost importance for electronic marketing. The number of page views per site indicates how satisfied users are with a given site or how useful a site is to them. If users in fact view many pages on a given site, it is likely that this site interests them. However, **if** individual users have the same "capacity constraints" that determine the extent of their Web utilization as measured by total number of pages viewed in a time period, then one could develop a theory based on P, the number of page views per site and N, the number of sites visited in the time period. Heavy users, as measured by the number of distinctive Web site visited, have on average a larger value for N implying that their value for P should be lower than the value of P for light users who visit few distinctive sites but visit each site with great intensity. To test this theory, we investigate whether the intensity of usage as measured in page views / site varies by trajectory group.

---

[24] While we believe panel tenure to be the most significant difference between our data, there are other differences that could potentially account for the different results. The Jupiter Media Metrix (JMM) panel is a national panel whereas the HomeNet panel is limited to the Pittsburgh area. Further, the data they use covers the period July 1997 to December 1999 while our data is from June 1995 to April 1997. Another important difference relates to the manner in which data has been collected in the Homenet study and the JMM panel. The JMM data is collected using a PC Meter resident on the machine being used by the panel member. In contrast, HomeNet data is collected using a proxy server. The meter permits fine-grained monitoring of user actions on the desktop. For example, the meter can record when a browser window is activated and the length of time that it is the active window. Proxy servers do not measure this sort of information. Thus, "Web viewings", the measure used by JMM is not the same as the measure of page views we use in this paper. A "Web view" begins when a Web page is accessed and the browser window used to access the page is active. If the browser window is de-activated, the Web view ends even if no new Web page has been accessed. In contrast a page view is defined by the user access to a Web page and is not related to window activations. To summarize, apart from differences in the process of collecting data, measuring Web use, and the nature of the panel, our results shed light on the steady state behavior of users while the work of Montgomery and Faloutsos focuses on the short term behavior of new users.

Tab. 9 reports descriptive statistics for the last 50% of the observation period by which time users appear to have reached a steady state of Web usage. Tab. 9 reveals that there are no material differences between trajectory groups in terms of their utilization intensity as measured in page views per site. Further, the theory that users with fewer distinct Web site visits have a higher number of page views per site is not supported by the data when users achieve steady state behavior. Moreover, it seems that no direct relationship between distinct sites visited and page views per site can be identified. When users achieve a steady state, light users are the least loyal group. This is contrary to the hypothesis that some individuals are visiting a large number of distinct sites infrequently and other users are visiting relatively few sites with high frequency. The issue of Web loyalty is discussed in more detail in chapter 5: 'Web User Loyalty and Web Site Stickiness' on pp. 80 ff.

Tab. 9: Summary information on page downloads by trajectory group (last 4 months only)

	light users	moderate users	heavy users	very heavy users
Average number of distinct sites visited per month	3.2	18.5	34.1	183.8
Average number of page views per month	10.8	92.3	153.8	817.5
Ratio of average monthly page views per distinct Web site	2.95	5.12	4.60	4.42

We also examined whether the intensity rates were an exaggerated summary statistic of the typical level of usage intensity due to users having a favorite portal they visit very often. The four most dominant portals / search engines in the data set were Infoseek (0.25% of the records in the set), Yahoo (1.99%), Lycos (0.30%), and Excite (0.70%). Together, these sites account for 3.24% of the page views.

Of specific interest was whether trajectory groups use portal sites and search engines differently. For example, users with low or moderate utilization rates may visit more 'one-stop-surfing' sites such as portals but use these sites more intensively than other users. Indeed, the analysis revealed that the percentage of these sites in the set of very heavy users is only 1.29%, whereas it is 4.00%, 4.77%, and 6.30% for heavy users, moderate users, and light users respectively.

While this difference is not quite significant at the .0f level (Prob > F = 0.0898), all the page requests of Web servers on the four most popular portals/search engines are removed from the data set: ‘yahoo.com’, ‘excite.com’, ‘lycos.com’, and ‘infoseek.com’. Deleting these records reduced the size of the page view data set by 3.24%. Tab. 10 reports the results of this analysis.

Tab. 10: Comparison of page views per site by user groups with and without search engines (last 4 months only)

	light users	moderate users	heavy users	very heavy users
Average of individual monthly page views per distinct Web site (with portals)	2.95	5.12	4.60	4.42
Average of individual monthly page views per distinct Web site (without portals)	2.89	4.90	5.15	4.36

Except for the group of heavy users, we see a slight decrease in the ratio of page views per site. Judging from these results, the advent of Web portals had only a minor effect on individual Web usage. The issue of portal utilization is discussed in more detail in chapter 6: ‘Portal Utilization’ on pp. 99 ff.

In summary, this study reveals that the population of residential Web usage can be clustered into four groups with distinct trajectories of use, whose usage behavior is not increasing but asymptotically saturating over time. Further, this finding does not seem to be dependent on the specific measure of number of distinct Web sites visited per month. An analysis of the distribution of page views over time leads to the same conclusion. Finally, the intensity with which people utilize their favorite Web site differs slightly across user groups. However, the average intensity rate did not seem to be skewed by portal sites. See chapter 6 ‘Portal Utilization’ on pp. 99 ff. for a detailed discussion of Web portal sites.

3.3.3 Group Profiles

In the next stage of the analysis, we examined the demographic profiles of the trajectory groups, thereby identifying characteristics that distinguished individuals following these four distinctive trajectories. To perform this analysis individuals were assigned to the trajectory group that best conformed to their actual usage trajectory. This assignment was based on the posterior probability of group membership. Based on the model's estimated parameters, it is possible to compute the probability of each individual's actual usage level over time, as measured by distinctive Web sites visited per month, conditional upon membership in a specific trajectory group [Nagin99a]. This probability is called the posterior probability of group membership. Individuals were assigned to the group with the largest such probability.

Tab. 11 shows the demographic differences across the groups with low and heavy Web usage and their statistical significance. Note that we aggregated the groups of light users and moderate users on the one hand, and the groups of heavy users and very heavy users on the other hand.

The summary statistics reveal that there is a difference in age across groups. Heavy users and very heavy users tend to be younger whereas light users and moderate users tend to be older. More significantly, there is a race effect and a gender effect. Individuals in the groups that use the Web heavily tend to be male and white. Conversely, the groups that make little use of the Web ('light users and moderate users') were disproportionately comprised of females and minorities. These results conform to the results reported in Kraut et al. [Kraut96a] and other studies on the digital divide. We also tested for the influence of household income on Web usage. Surprisingly, there is no income effect. We discuss the implications of this and the observed race and gender difference from the perspective of the digital divide in Section 3.4.3 on pp 60 ff.

Tab. 11: Overview of characteristics of users in the various groups

	all users	light users and moderate users	heavy and very heavy users	Prob > chi2 Prob > F
Percentage	100%	82.70%	17.30%	
Adult	74.30%	73.27%	78.89%	0.60
Female	51.40%	57.02%	26.81%	0.01
Minority	27.60%	31.35%	9.78%	0.05
Average age (years)	31.91	32.42	29.50	0.46
Household income	54,410	54,770	53,720	0.74
Computer skill[25]	3.43	3.33	3.89	0.03
Connection time (hours weekly)	2.15	1.65	4.08	0.00
Mail usage[26]	1.62	1.45	2.35	0.16
Phone usage[27]	4.14	4.14	4.25	0.87

E-mail usage and connection time were compared to other indicators of Internet usage. Not surprisingly, connection time increases as Web usage increases. However, as previously noted this measure of connection time is suspect. We studied e-mail usage and its relationship to Web usage because these are substantively different Internet services. The Web enables information retrieval. E-mail permits communication. Perhaps users who do not use the Web intensively may be using their time in communication activities. This is not supported by the data. E-mail usage actually increases with Web usage but the

[25] self reported 5-point scale ranging from 1 (low) to 5 (high)

[26] self-reported (0 = never, 1 = less than weekly, 2 = weekly, 3= few times/week, 4 =daily, 5=multiple times per day)

[27] self reported 5-point scale ranging from 1 (low) to 5 (high)

difference is not statistically significant across groups. Notice, however, Web-based e-mail, such as Hotmail or Yahoo! Mail, would be counted as a Web service vs. e-mail. Understanding the relative utilization of different Internet services is a topic of future research. See also [Kraut99] for further details on the relationship between Web usage, e-mail usage, and hours spent online by individuals.

Next, we tested if Internet usage is a substitute for telephone usage. For example, users might send e-mail to friends instead of calling them. Also, users might retrieve information from the Web instead of calling somebody to get the needed information. However, this does not seem to be the case. Phone usage actually increases slightly as Web usage increases.

3.4 Conclusions and Future Work

3.4.1 Major Results

This chapter explores whether the increase in Web site visiting opportunities spurred an increase in the Web utilization rates of individual users. A semi-parametric, group-based statistical method is applied to longitudinal data on individual Web usage.

The major results presented in this chapter are as follows:

- Web usage is not distributed equally across subgroups of users. Web users can be clustered into four groups with distinct trajectories of Web usage. The four identified groups of Web users are labeled ‘light users’ (52.5% of the population), ‘moderate users’ (30.2%), ‘heavy users’ (13.7%), and ‘very heavy users’ (3.6%).
- All groups reach saturation in their extent of Web usage after following a downward path. The saturation levels of these groups are 3 (‘light users’), 20 (‘moderate users’), 33 (‘heavy users’), and 200 (‘very heavy users’) distinct Web sites per month.
- We observe saturation of Web usage independent of the specific measure of distinct Web sites visited.
- Surprisingly, there are no material differences between trajectory groups in terms of their utilization intensity as measured in page views per site.
- Individual characteristics, particularly gender and ethnic background, determine the saturation levels of Web usage. Minorities and females utilize the Web to a lesser extent.

3.4.2 Implications for Electronic Commerce

The results presented in this chapter have important implications both for business-to-consumer electronic commerce and for public policy as it pertains to the digital divide. The Web can be thought of as a marketplace with sites competing to attract users to visit. The saturation levels for Web site visits in every trajectory group identified in this analysis can be interpreted to estimate the size of this market. For example, users visit on average about 33 distinct Web sites/month. If this were generalized to the online Web browsing population (let this number be N) at large, 33*N estimates the number of potential Web site visiting opportunities that Web sites will compete over each month. However, over the period of observation, the number of Web sites has grown exponentially. According to [IDS00] and [NUA], there were 72,398,092 sites and N=248,660,000 users online in January 2000. Thus, 72,398,092 sites are competing for these 33*248,660,000/month visiting opportunities. Over the past several years, N has continued to grow as the Internet has attracted new entrants. However, as the number of new entrants begins to decrease (this is already happening as can be seen from the estimates of online users at [NUA]), the number of Web site visiting opportunities will reach a steady state and we expect competition among Web sites for these visiting opportunities to grow in intensity.

The result of saturation is confirmed by a recent article by Nielson rating [NUA01], which reports that Web users have been spending less time online both at home and at work lately. The average time spent online fell by 15% between October and December 2001. The average number of online sessions per month dropped as well. The number of visits to unique sites fell from 20 to 17. This decline is consistent across all demographic groups. Also, the finding that Web utilization co-varies with time is also reported in a recent article by Horrigan & Rainie [Horrigan02]. They make the case that more recent adopters of the Internet are using it less intensively, which leads to a decline in a single person's history online.

On the individual level, the competition for Web users seems to be a zero sum game. On average, winning a user for one Web site means loosing this user to another site. Means of Web commercialization such as the advent of banner advertisements, which became common in the Web during the period of the study, does not seem to lead to an increase in Web usage in terms of visits to distinctive Web sites. The effectiveness of banner ads, which are supposed to trigger a higher number of visits to distinctive Web sites, might be questioned. This confirms existing sources of consumer response data on banner ads that propagate that with increasing exposure to passive banner ads, the probability that a consumer will click on it becomes close to zero [Double96].

Discussions of Web site visiting opportunities are relevant to business models in use in business-to-consumer electronic commerce. To date portals such as Yahoo! have relied almost exclusively on advertising income generated from serving banner advertisements (so called page impressions). This business model is dependent on maximizing visits from individuals – both first time and repeat visitors- in each time period. Portals have implemented a variety of personalized services (e.g., my.yahoo.com) to attract and retain visitors with varying degrees of success. Among the recent wave of dot com failures are several well funded portal sites. These include generic horizontal portals such as the Go portal (funded by Disney) and vertical portals such as drkoop.com that failed to garner sufficient Web site visitors to sustain themselves – a situation further exacerbated by the decline in online advertising and online advertising rates.

In contrast to portal sites, e-retail sites have to convert visitors into buyers, manage churn rates which represent loss of customers to the competition and enhance repeat purchase rates. Given limited capacity for Web utilization, sites that can achieve high rates of repeat visits and purchases are likely at a clear advantage. Supporting this hypothesis is a recent article by Agarwal et al. (2001) that reports on key processes in business-to-consumer commerce and states that successful retail commerce companies need to achieve visitor conversion rates of 12 percent, customer churn rates below 20 percent, and repeat purchase rates of around 60 percent [Agrawal01]. While these results do not shed light on the details of these conversion processes and how online companies should achieve these targets, limited capacity for Web site visiting opportunities among individuals is an important determiner of competition in this area.

This discussion highlights the need to understand the reasons underlying the capacity limits observed. It is possible that the limited capacity for Web site visits is due to the current technical shortcomings on the Internet (e.g., ease of use of sites, difficulty in using search engines, ineffectiveness of banner advertisements). Breakthroughs in technology can potentially increase the capacity for Web utilization and in turn the size of the market. For example, recent surveys such as [WSJ01] demonstrate that ads returned in response to searches are effective in increasing clickthrough rates. Similarly, recent studies undertaken by MSN, Cnet and Doubleclick also demonstrate improved effectiveness of online marketing campaigns using reengineered advertising technology [IAB]. However, it might well be the case that capacity limits on Web utilization are based on cognitive limits and cannot be mediated by technological breakthroughs [Pool84].

The increasing number of Web sites available in general and the increasing number of hits after querying a search engine in particular is in this regard what is commonly referred to as ‘information overload’. Information overload describes a situation when someone

finds an unmanageable amount of information, such as lots of emails in his inbox [Ganzel98]. Early research [Miller56] dealt with this issue in a general way. More recent research [Jones00a, Jones00b] links individual processing limits to online behavior, although it deals with computer mediated communication in a narrower sense by focusing on media that require a higher level of interaction, such as e-mail and IRC. Generally speaking, a way to deal with information overload is clearly to change or restrict behavior so that information becomes manageable, which indirectly anticipates the saturation found in the data. However, a deeper analysis of cognitive processing limits of individuals using the Web is beyond the scope of this dissertation and is a topic that requires further research.

Given cognitive processing limits of individuals, the finding of saturation in the number of Web sites visited on a monthly basis has potentially important implications for personalized Web search engines and personalized recommender systems. Given an individual 'capacity' for Web usage, such systems could help to find the best mix of Web sites. On the other hand, personalized search engines and recommender systems can trigger an increase in the number of Web sites visited by offering Web sites that better match the user's interests. In this regards, such systems help to make information manageable, even in a world of information overload.

3.4.3 Implications for Public Policy

Several of our results have a direct impact on the ongoing debate about a possible digital divide (see chapter 7: 'The Digital Divide Exists' on pp 111 ff.). As discussed in Section 3.3.3 on pp. 55 ff. and shown in Tab. 11, there are race and gender differences in the trajectory groups. For example, the percentage of people who belong to a minority group is 27.60% in the overall sample, 31.35% for light users and moderate users, and 9.78% for heavy users and very heavy users. Similar differences in the utilization of the Web by gender can be observed. For example, 51.4% of the people in the HomeNet sample are female. This percentage decreases as usage increases (moderate users: 58.1% female, heavy users: 28.6% female, very heavy users: 20.0% female). These findings imply that increased utilization of the Web will require more than access. As noted, all users in the HomeNet panel received free computers with Internet connections and basic training in use of the technology. The information on the Web is probably more appealing to whites than blacks. Therefore, more online information that is appealing to the black community is needed. Informal reports indicate that customized training by gender or race may be needed in addition to access to enable different segments of

society to benefit fully from the Internet. [Shade] proposes gender-sensitive training to meet the diverse needs of the female Internet user community even though recent studies such as [Cummings02] show that the gender gap is closing in terms of time spent online; men & women use the Internet different in terms of services used [Boneva01]. In this regard, additional work is required to develop policies that will be more successful in promoting utilization of the Web.

3.4.4 Future Work

This chapter contributes to the literature by presenting the results of a long-term study with residential subjects. Future work seems necessary with respect to three major issues: age of data, length of period of observation, and representativeness of the sample.

The patterns of Web usage found were based on usage data from 1995-1997. Technical advances, e.g., in the field human computer interaction in general or personalized recommender systems in particular can affect the intensity of Web usage (see [Spiekermann00, Görsch00] for a discussion of recommender systems). One of the main reasons for not using more recent data was to make sure that each user has a natural starting point of individual Web exposure. Further studies on people who did not use the Internet before are necessary to confirm the findings from 1998 onwards.

This study is distinctive in its use of 8 months of continuous individual Web usage data. While recent work by Montgomery and Faloutsos [Montgomery00] have used more recent data from the Jupiter Media Metrix Panel, the average tenure of users in their panel is much shorter and is biased towards behavior exhibited by new users in the short term. However, in order to gain insights in truly long-term changes in individual access behavior, the analysis of even longer samples of longitudinal data is desirable.

This study relies on the subject group of people from the HomeNet project, which is an opportunity sample, selected among families who were on boards of directors of neighborhood facilities or kids on school newspapers. Observed development in browsing behavior might arise due to cultural and social peculiarities of the subject group. Also, a significant share of the population accesses the Internet at work. Therefore, future research is necessary in order to confirm the findings for all groups of users. A truly random nationally representative sample is necessary for this work. Further, more work is necessary to confirm the results in international settings.

4 Analyzing Web Sessions

The results from the previous chapter: 'Saturation of Lay Web Usage' tell us that Web users in the HomeNet sample reach saturation in Web usage over time. This section aims at confirming this finding by analyzing the HomeNet data with more subtle, time-based measures. Like the previous chapter, changes in Web usage associated with increased experience of using the Web are detected. Specifically, it uses advanced measures that are based on user session to answer the question whether or not users shift from undirected browsing in the Web to directed access of Web sites as they gain expertise in using the Web. The results of this analysis have several important implications both for business-to-consumer electronic commerce and for public policy as it pertains to the digital divide.

4.1 Introduction and Motivation

A lot of research has been conducted on the relationship between an individual's use of information technology and his experience. Specifically, the level of expertise of a given user is supposed to affect usage of electronic information systems [Hsieh93, Fenichel81]. However, little is known about the extent to which individuals utilize the Web and gain expertise over time.

Using the same data from the HomeNet project, the previous chapter found that as individuals gain more Web browsing experience, the number of distinct Web sites the same individuals visit per month decreases to a saturation level that depends on whether the individual belongs to the group of light, moderate, heavy, or very heavy users. In this regard, the analysis in the previous chapter focused, like McKenzie et al. [McKenzie01a], on the changes in the user's vocabulary of Web sites over time. Figure 11 on page 48 depicts the results of that analysis. All users started using the Web in month 1. Thus, we consider the longitudinal development in Figure 11 on page 48 'learning curves' after a natural starting point of having first access to the Web.

However, the measure of distinct Web sites visited per month presented in the previous chapter is not entirely accurate in the sense that it does not take into account Web usage behavior within distinct Web sessions. Also, a decrease in the number of distinct sites visited per month may be due to a variety of reasons. For example, users may develop loyalty to Web sites and converge to a favorite set of sites. On the other hand, users may spend less time in the Internet. Thus, we wanted to measure time spent in the Web more

accurately using a session-based approach. Therefore, we followed the lead of Cooley [Cooley00] by dividing the individual clickstream into Web sessions. We computed five key measures of Web usage within Web sessions for each of the subgroups of users identified in the previous chapter.

The demographic factors that distinguish different user groups and the estimated proportion of the population belonging to each of these groups are identified in the previous chapter in Tab. 8 and Tab. 11 (see page 48 and page 56). The statistically significant determinants were age, gender, and race. In this chapter, the higher accuracy of measuring Web usage following a session-based approach helps to identify more determinants of Web usage.

The chapter is structured as follows: Section 4.2 describes our five key measures of Web usage in Web sessions. Section 4.3 presents the results of this analysis: the development of individual Web usage in Web sessions. Also, that section reports the results of a formal regression analysis, which reveal the individual characteristics (e.g., demographics of users) that determine Web usage in Web sessions. Finally, section 4.4.2 discusses the results and their implications for electronic commerce and public policy.

4.2 Measurements of Session-Based Web Usage

Like the previous chapter, this chapter adopts a longitudinal-development approach by monitoring the same individuals at different points in time. Over time, each individual gains experience and expertise (see Fidel et al. [Fidel98] and Khan et al. [Khan98] for the relationship between experience and expertise). In this longitudinal-developmental study, we took repeated measures of Web usage with respect to each individual.

As noted in section 3.2.1 on pp. 41 ff., there are several conceptually reasonable alternatives to take repeated measures of Web usage, which may be classified into frequency-based measures and time-based measures. Frequency-based measures such as the number of distinct Web sites visited by a given user per time period or Web session and total Web site visits per time period or Web session have been discussed in detail in the previous chapter.

Time-based measures use the time spent by an individual in the Web as an indicator of the utility received from utilizing the Web. Although the HomeNet project provides data on individual log-on and log-off activities, this time-based measure is not confident in the sense that it does not necessarily reflect time spent on the computer. Users may download a Web site but only actively attend to the Web site for a small fraction of time in

which the site was displayed (i.e., leave the computer unattended for hours or even days). Therefore, it was necessary to construct a more sophisticated measure that accurately reflects actual time spent actively interacting with a Web site. Thus, we followed the lead of Cooley [Cooley00] by compartmentalizing the individual clickstream into Web sessions. Catledge and Pitkow [Catledge95] published a study of Web browsing behavior and determined session boundaries by analyzing the time between each event for all events. Using a task-oriented method, they determined session boundaries by analyzing the time between each event for all events. According to that research, a lapse of 25.5 minutes or greater indicated the end of a "session." This heuristic is currently the most-commonly used for delimiting sessions (see also Choo et al. [Choo00]). We applied the same sessionizing criterion: A Web session is considered finished after a period of 25.5 minutes of user inactivity. Whenever a page request occurs more than 25.5 minutes after the previous request, it is considered the starting point of a new session. Cooley et al. [Cooley99] show that this approach leads to sufficient accuracy in identifying Web sessions. However, it represents a trade-off between accuracy in session identification and computational effort. Identifying user sessions remains a non-trivial problem. See Cooley [Cooley00] for a discussion of this issue.

We computed a variety of time-based and frequency-based metrics. In order to gain insights in the development of Web usage in Web sessions, we apply five key measures to the data from the HomeNet project. Specifically, frequency-based measures are combined with time-based sessionizing by tracking:

a) the development of the number of Web sessions over time,

b) the development of the number of distinct sites per Web session over time,

c) the development of the number of page views per Web session over time,

d) the development of the number of page views per Web site in Web sessions.

The measure d), ‘pages viewed per Web site within sessions’, deserves special attention because it speaks to the issue of Web loyalty, which is of utmost importance in electronic marketing. While the count of distinct Web sites visited provides an indication of individual-level willingness to search the exponentially expanding set of visiting opportunities, the number of pages viewed per distinct Web site permits an analysis of how intense users of these sites make use of these sites. In a case of complete loyalty where an individual’s page view capacity is directed to one Web site, this measure would be higher than the case where an individual page view capacity is directed to many sites.

We also applied the following time-based measure:

e) The development of the duration of individual Web sessions.

These measures are used to gain insights in how information searching behavior changes over time. Information searching behavior is supposed to change with increasing expertise in information technology [Hsieh93, Fenichel81]. Thus, we were interested in analyzing whether there actually is a trend to directed access of Web sites in contrast to undirected Web browsing. With respect to Web sessions, users may visit many distinct Web sites in the beginning of their Web learning experience when they start exploring the Web. The same users may visit less distinct Web sites per session later on as expertise in Web usage increases. It was reasonable to expect that less experienced users consume the Web in a few large chunks, whereas more experienced users consume Web sites in many small chunks. As expertise increases and users gain knowledge in using the Web, they may start a Web session for the very purpose of visiting one specific Web site (for example, checking a bank account using the Internet). Figure 14 illustrates such shift from undirected browsing to directed access of Web sites. The left hand of Figure 14 reports individual Web usage of a fictitious user, which consists of eight page views at three distinct Web sites: amazon.com, yahoo.com, and ft.com. Without applying a session-based approach we do not know whether the same user develops loyalty to Web sites. For example, on the left hand, the user shows limited loyalty to Web sites in the sense that he directs his Web usage to three Web sites, whereas the right hand of Figure 14 reveals that the same user shows complete loyalty to Web sites within each session. Therefore, in order to identify loyalty in sessions, we use the new measure of average page views per site in sessions. In this respect, users develop loyalty to Web sites if the ratio of pages viewed per Web site within sessions increases over time.

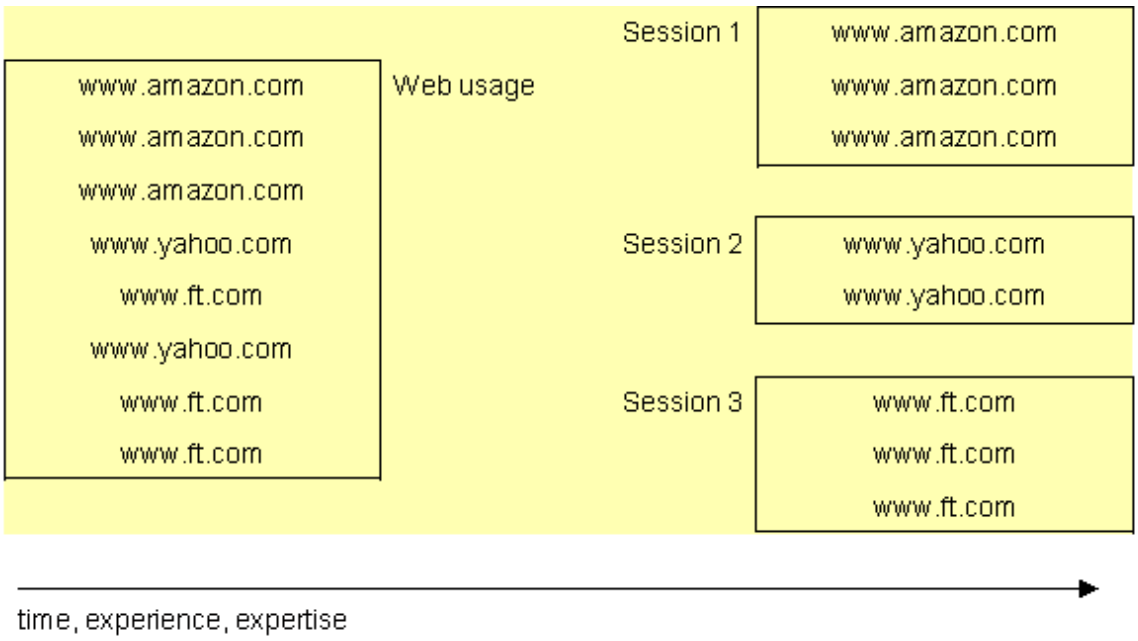


Figure 14: Hypothetical learning experience over time

It is reasonable to expect that a hypothetical learning experience such as demonstrated in Figure 14 has the following impact on the development of the measures a) to e). Specifically, given the example in Figure 14, it is reasonable to expect the following:

a) an increase in the number of Web sessions

b) a decrease in the number of distinct Web sites visited per session

c) a decrease in the number of Web pages viewed per session

d) an increase in the number of page views per Web site in Web sessions (an increase in loyalty to Web sites within sessions)

e) a decrease in the duration of Web sessions

We wanted to test if these expectations can actually be confirmed in the HomeNet data. Figure 14 also illustrates how aggregating Web usage in a given period of time (e.g., 1-month periods in the previous chapter) conceals patterns of Web usage in Web sessions. For example, suppose that the clickstream on the left hand of Figure 14 is the aggregated clickstream of a given user in a given month across all sessions. Suppose also that further analysis reveals that overall monthly Web usage can be divided into three distinct sessions. In each of these sessions the given user visits one distinct site only. Clearly, a measure such as ‘page views per site’ in a given month – as applied in section 3.3.2 on pp. 50 ff. - conceals loyalty in sessions. In this respect, this analysis advances the research on Web usage and Web loyalty presented in the previous chapter.

4.3 Results

4.3.1 Five Key Measures of Web Usage in Web Sessions

Following the sessionizing approach described above revealed that the 139 users in the HomeNet project use the Web in 6612 distinct Web sessions during the first 8 months of the project after their initial Web usage. The average duration of a session was 26.38 minutes. The average number of distinct sites visited per session was 6.39. Users visited on average 18.27 pages per session. Tab. 12 reports these descriptive statistics.

Tab. 12: Descriptive statistics on Web usage in Web sessions

total number of sessions	6612			
	Average	10th percentile	50th percentile	90th percentile
number of Web sessions	7.87	0	4	21
duration of Web session (minutes)	26.38	6	22	50
number of distinct sites visited per session	6.39	2.15	5.5	11.43
number of pages viewed per session	18.27	5	15	34.23
Ratio page views / site per session	3.03	2.5	2.57	4.33

Before moving on and analyzing how these measures evolve over time, it is important to notice there is a large diversity in the number of sessions, duration of sessions, and other key measures of Web usage across users. Because of this diversity, it is reasonable to observe sessionized Web usage for different groups of users separately, in order to answer the question if subgroups of users differ in how they divide overall Web usage into Web sessions. For example, do subgroups of users show different loyalty to Web sites in Web sessions? Do some users stay in the Web longer than other users? In this regard, we were interested in how the five key measures of sessionized Web usage differ across subgroups of people.

We computed the key measures for each of the subgroups of users identified in the previous chapter, in which users in the HomeNet sample were clustered into four distinct

groups: light users, moderate users, heavy users, and very heavy users, depending on how many distinct sites users visit per month.

Tab. 13 reveals that the number of sessions increased as overall Web usage increased. This relationship between overall Web usage and the number of Web sessions is almost tautological: users who use the Web rarely have fewer opportunities to visit Web sites. With respect to the duration of Web sessions, there is a substantial increase as overall Web usage increases. Note that the average duration of Web sessions of people who belong to the group of very heavy users was twice as long as the duration of sessions of light users. Therefore, heavy users not only log on to the Web more often, they also spent more time online and consume Web sites in larger chunks.

Tab. 13: Key measures of Web usage in Web sessions across subgroups of users – session count and session duration

	light users	moderate users	heavy users	very heavy users
Average number of sessions per month	2.6	7.6	13.3	36.1
Average duration of Web sessions (minutes)	21.0	26.8	32.5	40.7

Next, Tab. 14 compares the key measures: distinct sites visited, pages viewed, and the loyalty measure, ‘pages viewed per Web site’, using related monthly measures.

Tab. 14: Key measures of Web usage in Web sessions across subgroups of users – monthly vs. per-session metrics

		light users	moderate users	heavy users	very heavy users
monthly	Average number of distinct sites visited per month	3.2	18.5	34.1	183.8
	Average number of page views per month	10.8	92.3	153.8	817.5
	Ratio of average monthly page views per distinct Web site	3.0	5.1	4.6	4.4
Per Web session	Average number of distinct sites visited per session	5.3	5.9	8.1	11.5
	Average number of page views per session	14.3	17.6	24.4	31.9
	Ratio of average of monthly page views per distinct Web site per session	2.9	3.3	3.2	2.7

Notice that the number of distinct sites visited per month, which is the key measure used in the previous chapter as a basis for clustering four user groups, differ substantially. However, there is much less of a difference if we consider the number of distinct sites visited per session. For example, light users and moderate users visit almost the same number of distinct sites per session (5.3 and 5.9). The difference in monthly Web usage is apparently due to the lower number of Web sessions per month for light users rather than due to lower Web utilization in sessions. The same argument holds for the measures ‘average number of pages viewed per session’ and ‘pages viewed per Web site’.

A first measure of Web loyalty was introduced in section 3.3.2: 'Intensity of Web Utilization’ on pp. 50 ff., the number of monthly page views per Web site. Note that the advanced measure of Web loyalty used in this chapter, the ratio of page views per site in sessions reported in Tab. 14, is lower than the ratio of page views per site and month, as reported in Tab. 12. This is an indicator (though limited) of loyalty of Web users to Web sites. Keep in mind that Web usage within a given month usually consists of many distinct Web sessions. Under the assumption that users do not visit Web sites randomly and there is an overlap in the identity of Web sites visited across sessions, the vocabulary set of

Web sites visited by a given user does not increase by the same factor that the number of sessions increases. Thus, the ratio of page views per Web site and month is supposed to be larger than the same ratio per session if a user shows loyalty to Web sites at least to some extent. In the case of complete loyalty, in which user's keep returning to the same Web sites, the domain vocabulary would be a constant whereas the number of page views increases in a linear way. The ratio of page views per site and month would be the ratio pages viewed / distinct Web sites per session multiplied by the number of sessions for this user in the given month. For example, the monthly ratio of pages viewed per Web site of a moderate user who is completely loyal to Web sites would be 3.3x7.6=25.1. In Tab. 12, this ratio is 5.1. The comparison of these two measures reveals that there is only very limited loyalty in Web sites visited. This conforms to findings on limited loyalty presented in chapter 3. More subtle measures of Web loyalty are developed in chapter 5:'Web User Loyalty and Web Site Stickiness' on pp. 80 ff.

4.3.2 Results of the Longitudinal Analysis

Next, we are interested in how the five key measures evolve over time. Figure 15, Figure 16, Figure 17, and Figure 18 report the longitudinal development of these measures.

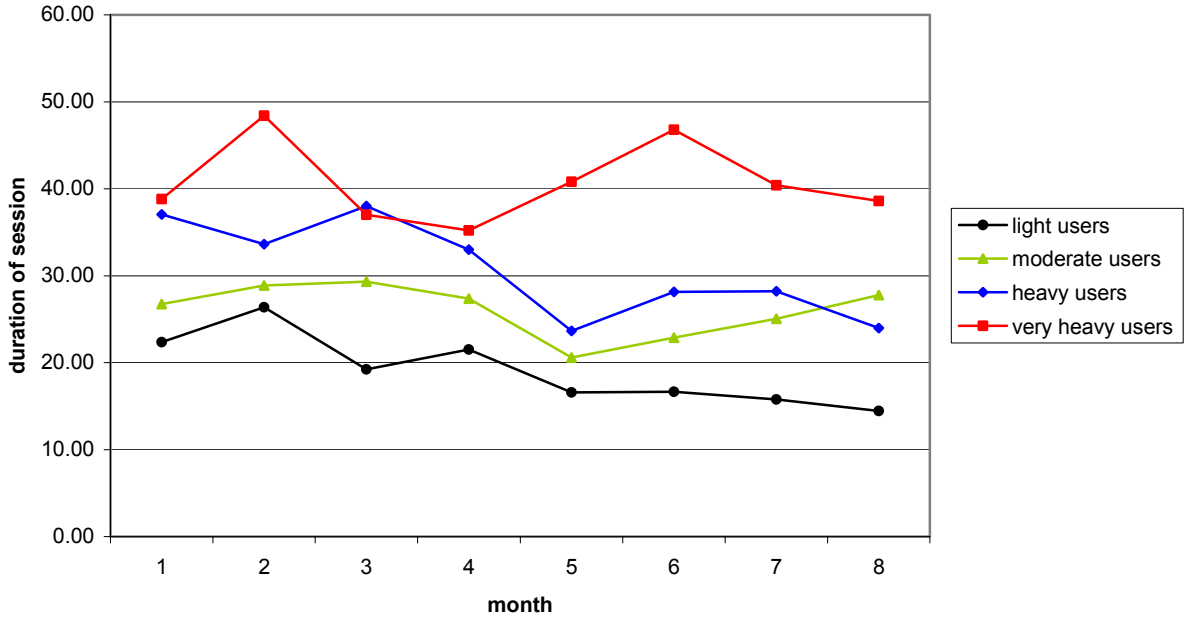


Figure 15: duration of Web sessions across subgroups of users

Figure 15 shows the duration of Web sessions across subgroups of users over time. Keep in mind that all users started using the Web in month 1. In this regard, we consider the longitudinal development of the key measures 'learning curves' after a natural starting point of having first access to the Web. Figure 15 reveals that the duration of Web sessions stays almost constant over time. There is only a slight decrease across subgroups of users. Most importantly, there is no group that follows an upward path. This seems to confirm the hypothesis that Web sessions become shorter as expertise of users increases.

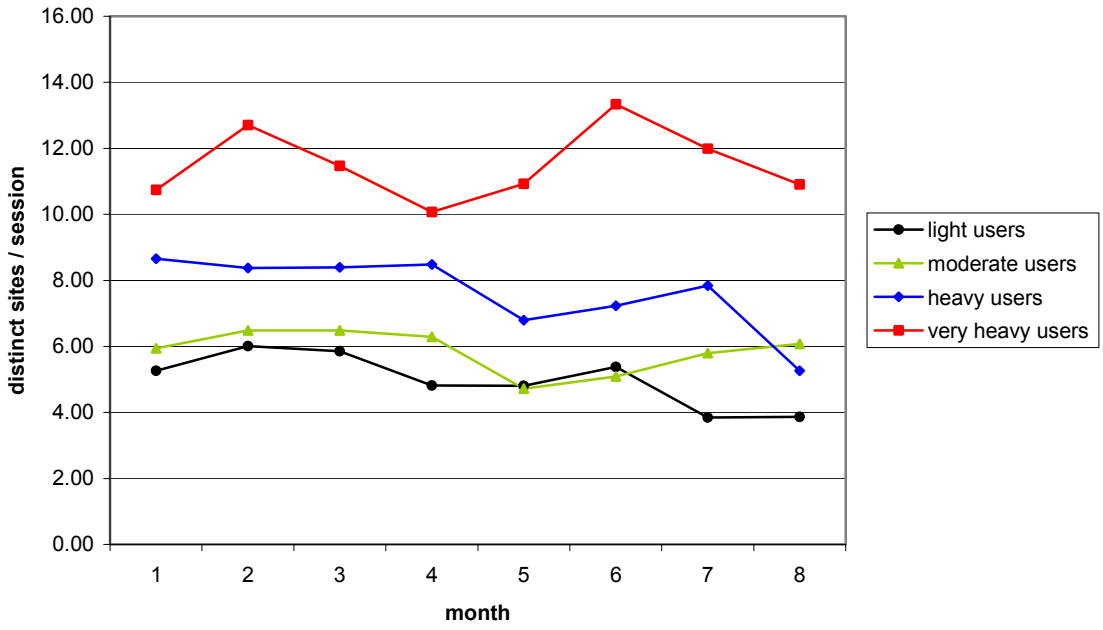


Figure 16: Distinct Web sites visited within sessions

Figure 16 reports the number of distinct sites visited per session by individuals. In contrast to the expectations stated in section 4.2, the number of Web sites visited within sessions did not decrease over time. In particular, light users visit nearly the same number of distinct sites per session as moderate users. In general, there is much less of a difference between the various groups than if measured in overall numbers. This seems to confirm that heavy overall Web usage is rather due to more frequent use of the Web than due to more intense sessions. The number of sites 'consumed' per session stays almost constant over time. For each subgroup of users, there is constant Web site consumption in sessions over time. There also seems to be a lower and upper bound for the number of distinct sites visited per session. Even heavy users do not visit more than 12 sites per session. On the other hand, even light users consumed quite a few Web sites in their Web

sessions. Overall low utilization rates for this group as described in the previous chapter are apparently due to a very limited number of Web sessions instead of limited Web usage within sessions.

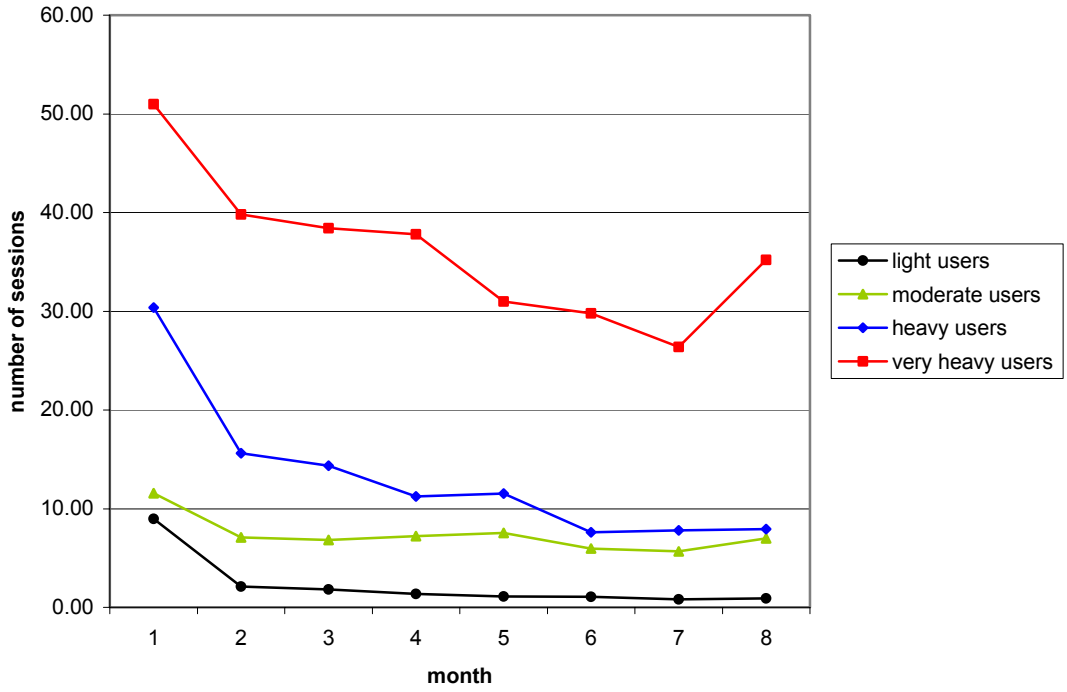


Figure 17: number of Web sessions over time

Figure 17 shows the number of sessions and how it evolves over time. Surprisingly, no group of users follows an upward trend. All groups follow a downward path until they reach saturation. Note that there is a large similarity between the trajectories depicted in Figure 11 on page 15 and the development of distinct Web sites accessed in Figure 17. This seems to confirm that overall Web usage is highly correlated with the frequency of Web usage.

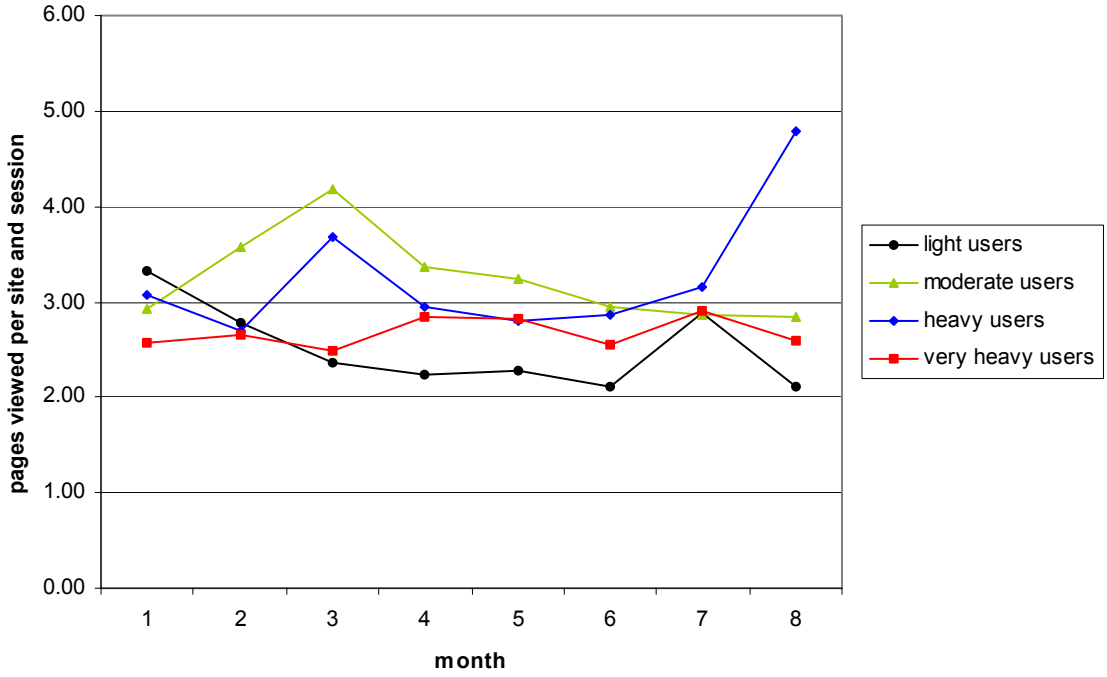


Figure 18: Pages viewed per Web site within Web sessions

Figure 18 depicts the development of the number of pages viewed per distinct site within Web sessions. Across subgroups, this measure stays more or less constant at a level of about three. Surprisingly, no group is following an upward path. In this regard, only very limited loyalty to Web sites in sessions develops over time. This conforms to McKenzie et al. [McKenzie01a], who find a lack of commonality in the sites visited by users (see also chapter 5:'Web User Loyalty and Web Site Stickiness' on pp. 80 ff.).

4.3.3 Results of the Regression Analysis

In order to determine the individual characteristics that determine Web usage within Web sessions, a more formal regression analysis is performed in this section. Specifically, we apply several POISSON models for which we chose the dependent variables 'number of sessions per month', 'number of distinct sites per session', 'page views per session', and 'duration of session'. Notice that applying models that use poisson distributions is very common when the dependent variable is a count. The regression tests which individual characteristics determine Web usage in Web sessions. The following variables are the parameters of these models:

- ‘white’, which tells something about the ethnic background of individuals (binary coded),
- ‘adult’ or ‘age’, which tells something about the age group of individuals (binary coded),
- ‘female’, the gender of individuals (binary coded),
- ‘income’, household income (in thousand US Dollars),
- ‘c-skills’, psychometric computer skill level (self reported 5-point scale),
- ‘mail’, mails sent weekly (self reported 5-point scale),
- ‘phone’, time spent using the phone (self reported 5-point scale).

We included mail usage in the analysis because it is another important indicator of Internet usage. Telephone usage was measured as well because it might be a substitute to Internet usage.

Tab. 15: Poisson estimates: Determinants of number of Web sessions

```
                                                  Prob > chi2     =     0.0000
Log likelihood =  -307.8775                       Pseudo R2       =     0.3426

------------------------------------------------------------------------------
# Websessions|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |    .283475   .1333055     2.13   0.033      .022201    .5447491
         age |   .0103011   .0034153     3.02   0.003     .0036073     .016995
      female |  -.7061442   .1065865    -6.63   0.000    -.9150499   -.4972385
      income |  -.0080047   .0019257    -4.16   0.000     -.011779   -.0042303
     c-skill |   .1535055   .0467751     3.28   0.001      .061828    .2451829
        mail |    .194439   .0136804    14.21   0.000     .1676259    .2212521
       phone |   .1080724   .0484168     2.23   0.026     .0131773    .2029676
       _cons |   .7840724   .3239362     2.42   0.016     .1491691    1.418976
------------------------------------------------------------------------------
```

Tab. 15 reveals that a white Caucasian ethnical background, age, computer skills, mail and phone usage correlate with the number of Web sessions. Being female and household income have a negative impact on this measure.

Tab. 16: Poisson estimates: Determinants of duration of sessions

```
                                                  Prob > chi2     =     0.0000
Log likelihood = -418.39166                       Pseudo R2       =     0.1206
------------------------------------------------------------------------------
    duration |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .1046968   .0602288     1.74   0.082    -.0133496    .2227431
         age |   .0042478   .0017762     2.39   0.017     .0007666     .007729
      female |  -.2456042    .053038    -4.63   0.000    -.3495567   -.1416517
      income |  -.0035302   .0010038    -3.52   0.000    -.0054976   -.0015629
     c-skill |   .1501078   .0254734     5.89   0.000      .100181    .2000347
        mail |  -.0397439   .0111477    -3.57   0.000     -.061593   -.0178948
       phone |   .0364762   .0264663     1.38   0.168    -.0153967    .0883491
       _cons |   2.745716   .1693715    16.21   0.000     2.413754    3.077678
------------------------------------------------------------------------------
```

The results of a regressions analysis of the dependent variable ‘duration of Web sessions’ show that white Caucasian ethnical background, age, computer skills, phone usage correlate with the duration of Web sessions. Being female, mail usage and household income have a negative impact on this measure.

Tab. 17: Poisson estimates: Determinants of number of sites per sessions

```
                                                  Prob > chi2     =     0.0005
Log likelihood = -206.61731                       Pseudo R2       =     0.0512
------------------------------------------------------------------------------
   Web sites |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .0720389   .1122298     0.64   0.521    -.1479274    .2920052
       adult |   .2545569   .1365446     1.86   0.062    -.0130657    .5221794
      female |   -.203511   .0970502    -2.10   0.036    -.3937258   -.0132961
     c-skill |   .1105222   .0506675     2.18   0.029     .0112157    .2098287
        mail |  -.0313086   .0229353    -1.37   0.172    -.0762609    .0136437
       _cons |   1.341345   .2492828     5.38   0.000       .85276     1.82993
------------------------------------------------------------------------------
```

Tab. 17 shows that a white Caucasian ethnical background, being adult, and computer skills have a positive impact on the number of Web sites per Web session. Being female and mail usage have a negative impact on this measure.

Tab. 18: Poisson estimates – Determinants of number of page views per session

```
                                                  Prob > chi2     =     0.0000
Log likelihood = -339.46156                       Pseudo R2       =     0.1669
------------------------------------------------------------------------------
pages viewed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .1564213   .0765005     2.04   0.041      .006483    .3063595
       adult |   .2176885   .0816161     2.67   0.008     .0577238    .3776532
      female |  -.3645287      .0638    -5.71   0.000    -.4895744   -.2394831
      income |  -.0018392   .0012002    -1.53   0.125    -.0041916    .0005132
     c-skill |   .1941714   .0305519     6.36   0.000     .1342907    .2540522
        mail |  -.0290614   .0137443    -2.11   0.034    -.0559996   -.0021231
       phone |   .0426757   .0318518     1.34   0.180    -.0197527    .1051042
       _cons |   2.069375   .1890624    10.95   0.000      1.69882    2.439931
------------------------------------------------------------------------------
```

The results of a regression analysis with the dependent variable ‘number of page views per session’ revealed that white Caucasian ethnical background, being adult, computer skills, and phone usage correlate with the number of pages viewed per Web session. Being female, mail and household income have a negative impact on this measure.

The implications of this analysis for electronic commerce and public policy are discussed in the next sections.

4.4 Conclusions and Future Work

4.4.1 Major Results

In this chapter, time-based measures are applied to the HomeNet data in order to advance the research from chapter 3. We introduced five key measures that help us to gain insights in individual Web usage in Web sessions. Moreover, we took repeated measures of Web usage in Web sessions over time in order to identify trends in Web usage. In this regard, we identified how Web users change the way they use the Web as their individual level of expertise increases.

The major results are as follows:

- Web users spent only limited time in the Web. Regardless of the way of measuring Web usage in Web sessions, only a small group of users uses the Web heavily.
- We identified characteristics of individuals that influence Web usage in Web sessions, which include ethnic background, gender, household income, phone usage, e-mail usage, and computer skill level. In particular, belonging to a minority group and being

female determines low Web usage in Web sessions, which confirms the findings from the previous chapter.

- Surprisingly, there does not seem to be a significant shift from undirected browsing to directed access of Web sites over time. This seems to confirm that users keep exploring the Web even after eight months of Internet experience, which is reasonable if one considers that the Web itself was growing tremendously over the period of observation. However, it was reasonable to expect at least the fraction of directed browsing to increase over time.

4.4.2 Implications for Electronic Commerce and Electronic Marketing

The results of this study have several important implications both for business-to-consumer electronic commerce and for public policy. The rejection of the hypothesis of increasing Web usage as described in chapter 3: 'Saturation of Lay Web Usage' on pp. 40 ff. is a first indicator that competition among Web companies for Web market share is likely to become more intense when the growth in terms of numbers of people accessing the Web slows down. We see this to a similar extent if we consider Web sessions. Further, judging from the results that users do not develop significant loyalty in the Web, it is rather difficult for Web site operators to retain customers.

The finding that overall Web usage highly correlates with the number of sessions could simply mean that residential Web usage depends heavily on the time available for Internet use at home. In fact, Figure 19 and Figure 20 reveal that usage varies considerably by time of day and day of the week. Web usage of light and moderate users becomes substantial in the later afternoon (probably after work), whereas heavy and very heavy users use the Web also during the day. On weekends, usage is spread more evenly. However, the prime-time hours in the evening remain the hours of heavy usage. Such information on when individuals use the Web is also crucial for planning updates of the site, promotion campaigns, and advertising.

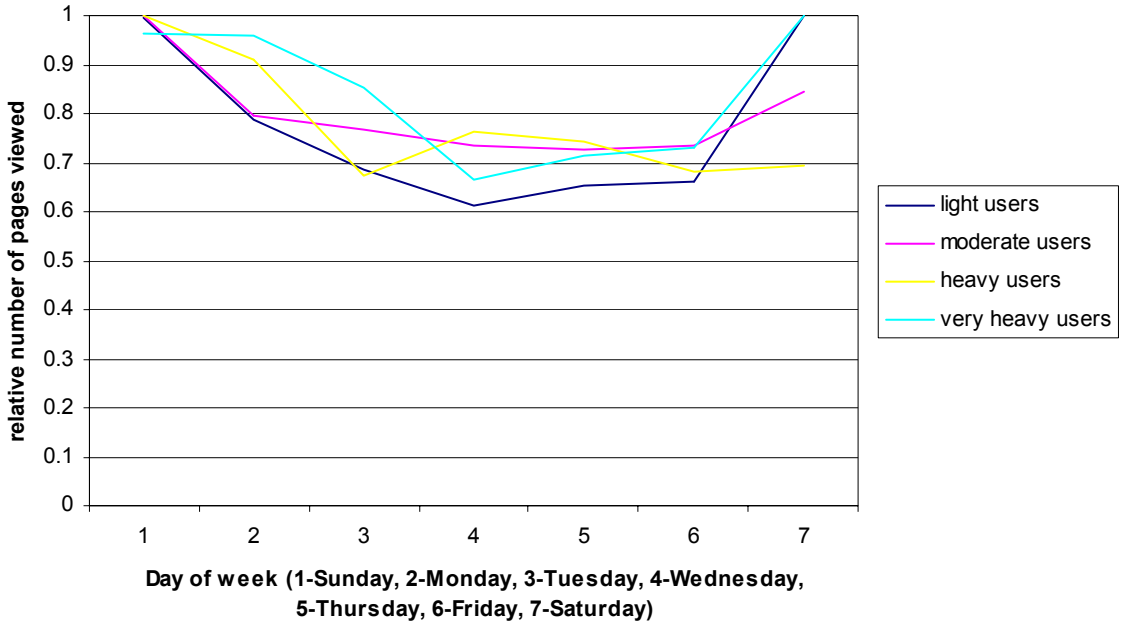


Figure 19: Distribution of Web usage by day of week across groups

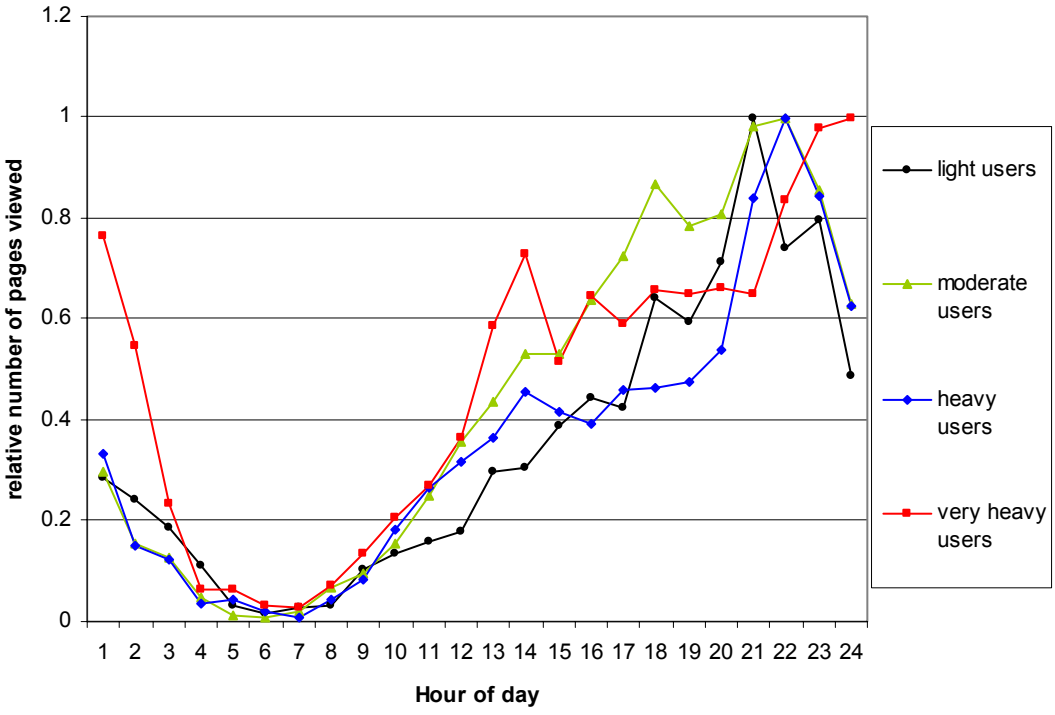


Figure 20: Distribution of Web Usage by hour of day across groups

This analysis also reveals that Web users seem to have a limited capacity of Web usage in Web sessions. For example, even the group of very heavy users used the Web on average for only 40 minutes. This is twice the duration the group of light users used the Web. In this regard, very heavy users not only used the Web more often but also stayed online longer. However, overall Web usage was determined by the number of sessions rather than by the duration of Web sessions. It may be that there are cognitive processing limits that are responsible for this [Miller56]. There seem to be capacity constraints not only with respect to the duration of Web sessions but also with respect to the number of pages viewed and the number of distinct sites visited per Web session. This was already anticipated in chapter 3. The findings on limited capacity with respect to Web usage highlight the need to understand the reasons that underlie these capacity limits observed. It is possible that the limited capacity for Web site visits is due to the current technical shortcomings on the Internet.

4.4.3 Implications for Public Policy

Using an approach that is based on Web sessions advances the research on Web usage and helps to identify determinants of Web use that otherwise would not have been identified. This analysis reveals that individual characteristics such as gender (male) and ethnic background (white) determine heavy Web usage in Web sessions significantly. Note that age has a positive impact on Web usage in sessions. Prior work on the HomeNet data such as described in chapter 3 did find a negative impact of age on Web use. Using a more subtle session-based approach of measuring usage has advanced this work and lead to new conclusions.

Household income does have an effect on Web utilization in sessions. In contrast to other studies on the digital divide, such as Kraut et al. [Kraut96b], high household income has a negative impact on Web utilization. Supposing that income correlates with time spent at the workplace, this may be due to the lack of time available for residential Web usage.

4.4.4 Future Work

Determining the reasons underlying the presumed capacity constraints is an important topic for future research. Also, because residential Internet usage seems to be highly dependent on the time available for Internet use at home, it is important to conduct similar studies at the workplace.

5 Web User Loyalty and Web Site Stickiness

Terms such as “churn” and “stickiness” have been used to describe loyalty on the Web. This chapter advances the work of the previous chapters by measuring churn of Web users and stickiness of Web sites in the HomeNet data sample. It is desirable to measure loyalty of users in the different trajectory groups to the Web sites they visit more accurately than by average number of page views / site (per session), as conducted in chapter 3 and chapter 4. This chapter introduces precise ways of measuring loyalty on the Web and characterizes loyalty empirically using the HomeNet data. The results have important implications for Web site operators from a business perspective.

5.1 Introduction and Motivation

The rapid growth of the Web depicted in Figure 1 on page 15 has led to many new research questions, such as the issue of Web user loyalty. Web user loyalty is a behavior that deals with user intentions to repeat visits at Web sites. In electronic commerce with online vendors, it is the foundation for customer loyalty, which is a behavior that deals with customer intentions to do repeated business with the vendor and to recommend this vendor to other customers [Zeithaml96]. In this regard, returning to a vendor’s Web site – the very act of visiting the Web site – is the precondition for customer loyalty online. Chow et al. [Chow97] and Heskett et al. [Heskett94] show that customer loyalty increases profit and growth of brick-and-mortar companies. For example, depending on the industry involved, increasing the percentage of loyal customers by as little as 5% can increase profitability by as much as 30% to 85% [Reichheld90]. There are several reasons for that: For example, loyal customers are typically willing to pay a higher price and are more understanding when something goes wrong [Chow97, Fukuyama95, Reichheld90, Reichheld00, Zeithaml96], and are easier to satisfy because the vendor knows the customer’s expectations better [Heskett94, Reichheld90, Zeithaml96].

This effect of increasing profitabily is estimated to be even stronger on the Web [Reichheld00]. The success of some well-known websites is inextricably linked to their ability to maintain a high degree of customer loyalty. For example, Amazon’s success can be attributed to its loyal customer base: 66% of the purchases are made by returning customers [Economist00]. The importance of customer loyalty in e-commerce is heightened by the high cost of attracting new customers on the Internet and the relative difficulty in retaining them [Gefen02]. Moreover, attracting new customers cost online

vendors at least 20%-40% than it cost traditional vendors. It often takes over a year of repeat purchases to recoup the initial cost of attracting Web users to a specific commercial Web site [Reichheld00].

If one considers the Web as a marketplace, the number of Web users multiplied by each individual's saturation level of Web usage (as identified in chapter 3 'Saturation of Lay Web Usage' on pp. 40 ff.: 3, 20, 33, and 200 sites per month for the group of 'light users', 'moderate users', 'heavy users', and 'very heavy users' respectively, see also Figure 11 and Figure 12 on page 48f) determines the size of the market. Clearly, the size of the market affects the nature of competition. Therefore, we concluded in chapter 3 that the Web is a highly competitive entity whose degree of competition is likely to become even higher when eventually the growth of the Web in numbers of new users accessing the Web slows down. Thus, the results from the previous chapters emphasize that research on the dynamics of usage has to incorporate an analysis of churn in the Web, in order to analyze whether a given level of Web usage intensity is directed to one site or many sites. There is the need to further analyze the extent of loyalty of individual users to Web sites.

Such analysis should identify the demographic characteristics of loyal and disloyal user groups and the identity and characteristics of the Web sites that engender the most loyalty. The demographic factors that distinguish different user groups and the estimated proportion of the population belonging to each of these groups are reported in Tab. 8 on page 48 and Tab. 11 on page 56. This chapter answers the question if these groups also differ in loyalty to Web sites. For example, it is reasonable to expect that users with lower Web utilization rates are more loyal to Web sites than heavy users, because heavy users may visit a large number of distinct sites infrequently and moderate users maybe visiting relatively few sites with high frequency.

This chapter is organized as follows: Section 5.2 introduces precise quantitative ways of measuring the loyalty of Web users to Web sites over time. It analyzes whether a given level of Web usage intensity is directed to one site or many sites. It thereby answers the question if users converge over time to a set of 'favorite' Web sites. Section 5.3 paves the way for measuring popularity of Web sites, which influences the probability that a given Web site will be in a user's set of favorite sites. Section 5.4 measures the 'stickiness,' which determines the ability of these sites to remain in the set of favorite domains over time, of the most popular Web sites. Section 5.5 brings together the results in a 'popularity-stickiness map'. Section 5.6 describes possible dynamics of Web site popularity and Web site stickiness. Implications for electronic commerce are discussed in section 5.7.2. Finally, section 5.7 deals with open research issues.

5.2 Churn in Web Sites Visited

The results from the previous chapters tell us that different groups of people reach different levels of saturation in terms of how many distinctive Web sites they visit over time. These saturation levels differ across groups. However, it is important to keep in mind that individuals do not necessarily visit the same distinct Web sites every month. Indeed, there might be considerable churn in the specific Web sites visited over time. Therefore, we are interested in the loyalty of users in the various groups[28] to the Web sites they visit. By measuring the degree of loyalty of Web users to Web sites over time and analyzing whether a given level of Web usage intensity is directed to one site or many sites, one could answer the related question about the demographics of the loyal users on the Web. In case of low loyalty or high churn, there would be limited overlap over time in the identities in the specific Web sites visited. When it comes to measuring churn over time, two extreme scenarios are possible:

- No churn

  When people reach saturation, they visit the same set of 20, 33, or 200 specific Web sites (for moderate, heavy, and very heavy users respectively – see section 3.3.1: 'Trajectories of Usage' on pp. 47 ff.) over each time period (e.g., every month).

- 100% Churn

  When people reach saturation, they visit 20, 33, or 200 Web sites (depending on group membership) per month but do not visit the same sites from one month to another.

In a 'no churn' scenario, people find their favorite set of Web sites which they stick to after a period of 'exploring' the Web. It would be very easy to detect the successful Web sites that 'survived' the exploration period of a given user by simply identifying the Web sites that remain in the user's set in the last period of observation.

However, it is reasonable to presume that the truth lies somewhere between the two extremes. Therefore, it is important to find the right measurement of churn over time, which involves a variety of issues. The fact that there may be sites to which users are loyal to should increase the measurement of loyalty of a given user. On the other hand,

---

[28] Analyzed were the groups of 'moderate users', 'heavy users', and 'very heavy users'. The group of 'light users' was excluded from the analysis of Web loyalty.

the fact that there may be sites to which users are not loyal to should decrease a measurement of loyalty. There are already some existing approaches of measuring churn in the Web related literature. For example, Tauscher et al. [Tauscher97a] simply use the percentage of revisits to Web pages over time. The previous chapters used the page view per site ratio. Now, a more sophisticated approach is proposed.

In the example in Tab. 19, a given user visits 4 distinct Web sites {A,B,C,D} in the first time period *t*=1, 3 distinct Web sites {A,B,D} in the second time period *t*=2, and three distinct Web sites {A,F,G} in the third time period *t*=3. Apparently, this user is loyal to Web site A, which he visited in all of the three periods of time. On the other hand, Web site C was only visited once in *t*=1 but not in *t*=2 and *t*=3, indicating disloyalty to this Web site. Tab. 20 depicts the case of complete loyalty. Every Web site is revisited in the period of time followed by period in which the site first appeared. On the other hand, Tab. 21 shows the case of total disloyalty, where there are no revisits at all.

Tab. 19: An example of sets of Web sites visited

t=1	t=2	t=3
A	A	A
B	B	F
C	D	G
D		

Tab. 20: Sets of Web sites visited - The case of complete loyalty

t=1	t=2	t=3
A	A	A
B	B	B
C	C	C

Tab. 21: Sets of Web sites visited - The case of complete churn

t=1	t=2	t=3
A	D	G
B	E	H
C	F	I

We apply the following method of measuring churn for each given user:

$$c_{i,t} = \frac{S_{i,t...t+T}}{\sum_{time=t}^{t+T} S_{i,time}},$$

where $c_{i,t}$ is the churn of a given user i in a given period of time t, the numerator is the number of Web sites visited by the same user in a time window that starts at t and ends at $t+T$, and the denominator is the sum of numbers of visits to distinct Web sites in the periods of observation t, $t+1$, …, $t+T$, of which the time window $t…t+T$ is comprised of. T is the fixed length of this time window. For example, T in Tab. 19 equals 3, the numerator is 6 (distinct Web sites), and the denominator is 4+3+3=10.

Applying this measure to the examples in Tab. 19, Tab. 20, and Tab. 21 leads to the following results: In Tab. 21 – the case of total disloyalty – the churn for a given user 1 is:

$$c_{1,1} = \frac{9}{3+3+3} = 1.$$

In Tab. 20 – the case of total loyalty – the churn for the given user 1 is:

$$c_{1,1} = \frac{3}{3+3+3} = \frac{1}{3}.$$

The churn for the given user 1 in Tab. 19 is:

$$c_{1,1} = \frac{6}{4+3+3} = 0.6.$$

Note that the upper bound of c is 1 and the lower bound of c is 1 divided by the length of the time window, which is 1/3 in the example. In other words: $(1/T) \leq c_{i,t} \leq 1$, where $c_t=1$ for the least loyal user and $c_t=1/T$ for the most loyal user.

We define $cn_{i,t}$ as the normalized measurement of churn with $0 \leq cn_{i,t} \leq 1$:

$$cn_{i,t} = 1 - \left(\frac{1}{1 - T^{-1}} \times \left(1 - c_{i,t}\right) \right)$$

In example A, B, and C, given a time window of T=3, $cn_{i,t}$ equals 0.4, 0, and 1 respectively.

The time windows with T=3 or any other length can be used as a sliding time window to capture the development of churn over time. We compartmentalized data in the HomeNet sample by using 1-month periods. We also used a sliding time window with an arbitrary chosen length of T=3 (months) to identify trends in Web loyalty.

Figure 21 depicts the average normalized churn of given groups of users in the HomeNet sample over a period of 8 months.

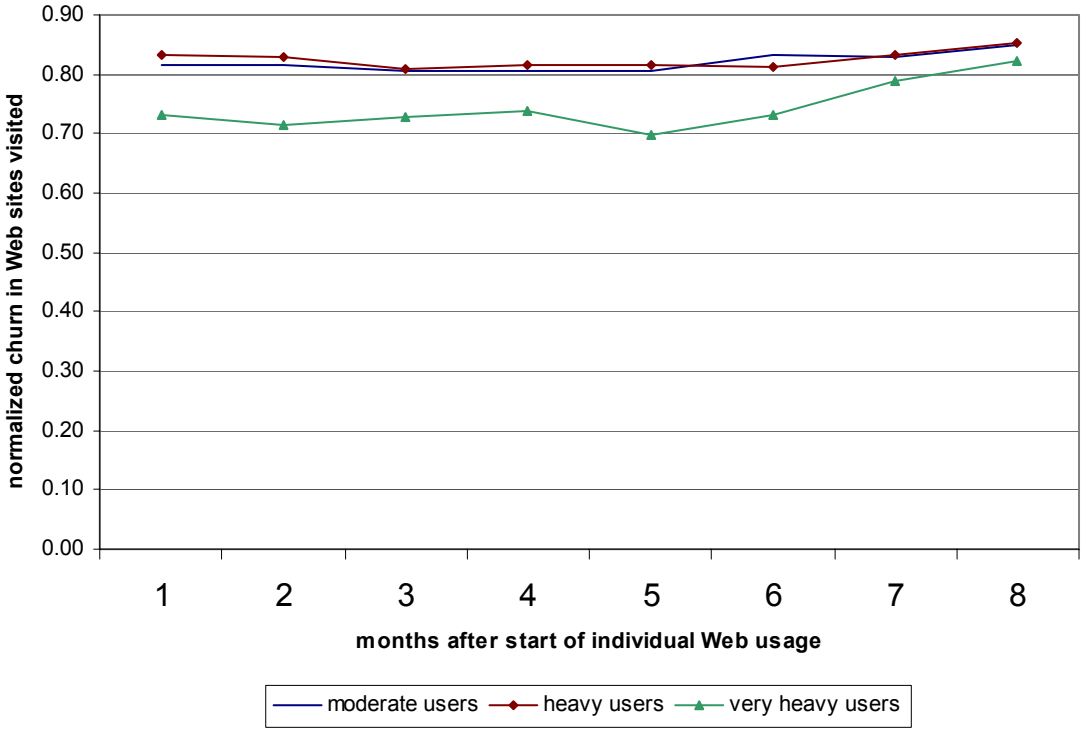


Figure 21: Normalized churn in Web sites visited by users in the HomeNet sample

While there is considerable churn across all groups in the sample, the surprising result is that over time, churn stays almost constant around 0.8 and is independent of group membership.

The results imply that moderate users who visit about 20 sites per month are loyal to about 4 in 20 sites. Heavy users are loyal to about 7 sites out of 33 sites they visit each week and very heavy users are loyal to about 40 sites out of the 200 they visit each month. Because churn is almost equal for all groups, group membership seems to have only a negligible impact on churn. It was reasonable to expect that moderate users would show a lower degree of churn. However, the findings contradict this hypothesis. The group of very heavy users seems to be even a little more loyal than the other two groups. In general, high churn or little loyalty, as shown in Figure 21, indicates that the process of exploring the Web continues even in late periods of observation. Because there is no trend of decreasing churn, users seem not to converge to a favorite set of Web sites. Even though there may be a set of favorite Web sites for given users, their percentage in the set of visited Web sites must be rather small. The use of these Web sites does not lead to a decrease in the number of new Web sites visited over time, which results in almost constant churn over time. Given the fact that the Web itself is a dynamic entity that offers exponentially growing visiting opportunities (see Figure 1 on page 15) this is a reasonable result.

Fortunately for Web site operators, a churn of about 0.8 means a loyalty of 0.2. As long as loyalty is not zero, users do not visit Web sites at random. Moreover, even if there are apparently only minor differences between the different groups of people percentagewise, there are substantial differences in the actual number of Web sites people are loyal to. For example, according to the results from chapter 3, the group of heavy users visits considerably more distinct sites per month than the group of moderate users. Both groups do have – percentagewise – about the same degree of disloyalty of 0.8. Given a loyalty of approximately 0.2 for both groups, the number of Web sites that users are loyal to is much larger for the group of very heavy users than for the group of moderate users. In general, there are - to a different extent depending on group membership - Web sites that are less affected by churn and thus are apt to stay permanently in the set of Web sites. Ways to identify these successful sites will be discussed in the following sections.

5.3 Popularity of Web Sites

As a measure of popularity of Web sites we organized the data by the number of users who accessed a given Web site in a long period of observations, such as 8 months, which is the whole period of observation in the HomeNet data sample. In this regard, popularity means the short-term popularity that derives from attracting users at least once without saying anything about the ability of the site to make users visit the same site again. Tab.

22 shows the most popular Web sites in the HomeNet sample, ordered by the percentage of users accessing each site.

Tab. 22: Most popular sites in the HomeNet sample

domain	overall popularity
homenet.andrew.cmu.edu	0.96
home.netscape.com	0.93
yahoo.com	0.79
cs.cmu.edu	0.59
excite.com	0.54
info.cern.ch	0.49
pathfinder.com	0.47
infoseek.com	0.45
pitt.edu	0.37
lycos.com	0.37
w3.org	0.34
mit.edu	0.29
pittsburgh.net	0.28

Judging from the results reported in Tab. 22, only a few Web sites were popular with most participants. This conforms with Adamic et al. [Adamic01] who report that 'millions of users flock to a few select sites, paying little attention to millions of others'.[29] The most frequently accessed services were the homepage of the HomeNet project and the Web site of Netscape. Also, among the most popular sites were directories, indexes, and portals, such as Yahoo and Lycos, which help people to find information either by providing taxonomies and search capabilities or by aggregating content.

Note that domains of banner ad sites and Web hosts have been removed from Tab. 22 because they skewed the results. Furthermore, note that the data show particular characteristics of the HomeNet sample. Users in this sample are people from the Pittsburgh area, which explains the high popularity of some local Web sites (e.g.,

[29] This can be expressed as a power law distribution [Adamic01]

pittsburgh.net). This also supports the hypothesis that a large share of Web activity is 'local', even if the Internet itself is 'global'. This is not necessarily surprising, since local information has special appeal to participants for a variety of reasons. For example, movie listings, bus schedules, and information on sport teams are more useful when they are local [Kraut96a].

5.4 Stickiness of Web Sites

The popularity of a Web site does not say much about the actual 'stickiness' of the same site, its ability to attract users again. Popularity of a Web site as measured in the previous section could be the result of many users visiting this Web site only once without ever coming back.

Therefore, more subtle measures of a Web site's success are needed. In this regard, we calculate the 'stickiness' of a given Web site as:

$$s_{i,url} = \frac{\#a_{i,domain}}{\#p_{i,domain}},$$

where $s_{i,domain}$ is the stickiness of a Web site *domain* for a given user *i*, $\#p_{i,domain}$ is the number of months left in the sample period after user *i* accessed site domain first, and $\#a_{i,domain}$ is the number of months after the user accessed the Web site first in which the users actually accessed the given Web site.

For example, Tab. 23 depicts the stickiness data of yahoo.com for a subset of users. Zeros denote months in which a given user did not access this Web site. 'Ones' denote months in which the user actually accessed the site. Missing data is denoted by dots. In this example, user i=52 accessed the Web site domain='yahoo.com' first in period 2. After that there remain $\#p_{i,domain}$=12 periods of observation (t3-t14). User 52 accessed the given site in $\#a_{i,domain}$=4 of the remaining 12 periods of time (namely in t4, t7, t8, and t13). Stickiness is calculated as $s_{i,domain}$=4/12=0.33.

Tab. 23: Stickiness table for Yahoo!

user	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	**Stickiness**
…															
50	1	1	0	0	1	1	1	1	0	1	0	0	0	1	**0.54**
51	1	1	1	1	0	1	1	1	1	1	1	1	1	1	**0.92**
52	0	1	0	1	0	0	1	1	0	0	0	0	1	0	**0.33**
53	1	1	0	0	0	0	0	1	1	0	1	1	1	0	**0.46**
54	0	0	0	0	0	1	0	0	0	0	0	0	.	.	**0.00**
55	0	0	0	0	0	0	1	0	0	0	0	0	.	.	**0.00**
56	1	0	1	0	1	1	1	0	0	0	0	0	0	.	**0.33**
…															
avg	**0.58**	**0.39**	**0.23**	**0.27**	**0.31**	**0.30**	**0.31**	**0.25**	**0.23**	**0.26**	**0.30**	**0.25**	**0.25**	**0.23**	**0.46**

The share of the users who actually accessed this Web site in a given month is shown in the line at the bottom of Tab. 23 (e.g., there is an average popularity of 0.23 in period t3). Furthermore, the number in the lower right corner of Tab. 23 shows the average stickiness of ‘yahoo.com’ across all users in the sample (0.46). Users who did not visit the Web site at all were dropped from the calculation of this average stickiness (missing data).

We created tables similar to Tab. 23 for the more popular Web sites given in Tab. 22. We focused on the popular sites because the smaller the number of users a Web site attracts, the sparser the data in such a table becomes, which complicates the accurate calculation of a Web site’s stickiness. Tab. 24 depicts the summary lines of all tables similar to Tab. 23 for the other popular Web sites in the HomeNet sample. Note that the lower the stickiness, the higher the difference between overall popularity as reported in Tab. 22 and average popularity over time.

Tab. 24: Stickiness and popularity of Web sites in the HomeNet data

domain	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	Sticki-ness	Overall popularity	Average popularity
HOMENET.ANDREW.CMU.EDU	0.94	0.69	0.66	0.68	0.59	0.65	0.64	0.57	0.54	0.50	0.61	0.63	0.61	0.69	0.71	0.96	0.64
HOME.NETSCAPE.COM	0.90	0.74	0.63	0.71	0.65	0.67	0.60	0.56	0.57	0.46	0.61	0.58	0.53	0.56	0.73	0.93	0.63
YAHOO.COM	0.58	0.39	0.23	0.27	0.31	0.30	0.32	0.25	0.23	0.26	0.30	0.25	0.25	0.23	0.46	0.79	0.30
CS.CMU.EDU	0.38	0.30	0.22	0.24	0.10	0.15	0.13	0.11	0.08	0.05	0.14	0.04	0.08	0.12	0.39	0.59	0.15
EXCITE.COM	0.27	0.25	0.16	0.17	0.11	0.08	0.12	0.07	0.10	0.05	0.08	0.08	0.05	0.04	0.33	0.54	0.12
INFO.CERN.CH	0.28	0.13	0.11	0.11	0.04	0.05	0.07	0.08	0.06	0.07	0.06	0.03	0.03	0.08	0.32	0.49	0.09
PATHFINDER.COM	0.25	0.20	0.14	0.13	0.04	0.07	0.08	0.07	0.03	0.05	0.06	0.04	0.03	0.04	0.30	0.47	0.09
INFOSEEK.COM	0.35	0.08	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.05	0.10	0.19	0.45	0.05
PITT.EDU	0.15	0.09	0.07	0.11	0.12	0.08	0.14	0.14	0.09	0.10	0.11	0.04	0.14	0.12	0.42	0.37	0.11
LYCOS.COM	0.16	0.04	0.03	0.04	0.02	0.02	0.02	0.06	0.05	0.00	0.10	0.03	0.12	0.12	0.25	0.37	0.06
W3.ORG	0.20	0.09	0.08	0.07	0.04	0.03	0.03	0.03	0.02	0.04	0.03	0.01	0.03	0.00	0.28	0.34	0.05
MIT.EDU	0.15	0.09	0.07	0.05	0.05	0.04	0.01	0.07	0.00	0.01	0.03	0.01	0.02	0.02	0.29	0.29	0.04
PITTSBURGH.NET	0.14	0.08	0.10	0.08	0.03	0.02	0.04	0.02	0.01	0.01	0.10	0.01	0.05	0.04	0.31	0.28	0.05

One might think that popularity and stickiness measure the same latent construct. This is not necessarily the case, especially because we measure the overall stickiness of a Web site only as an average of individual stickiness of the users who actually access the site. Even though the two measures are related, a popular site is not necessarily sticky and a sticky site is not necessarily popular. However, note that – on average – there is a relation between popularity and stickiness if you observe a fixed number of users over time. Notice that the data from the HomeNet project follows the browsing behavior of each user over the entire period of observation. Given a fixed set of users, even most popular Web sites will loose their popularity over time if they do not attract users again. In this regard, note also that local Web sites such as the Web sites of one of Pittsburgh's universities - pitt.edu - sustain a constant popularity over time.

Figure 22 displays the development of popularity of some Web sites over time. All Web sites in Figure 22 suffer a loss of popularity after the start of the project. However, some Web sites such as Yahoo! manage to sustain a quite constant popularity over time, which indicates that they have properties that make users come back to the site. In this regard, note that yahoo.com has a very high stickiness in Tab. 24 and a high overall popularity in Tab. 22. In a world with a fixed set of users (139 users in the HomeNet project), high stickiness and a high initial degree of popularity leads to constantly high levels of popularity. Another explanation for this is the word of mouth: if people have reasons to stick to a Web site (high stickiness), they might tell their friends to visit this site, thereby increasing the number of users at this site (high popularity).

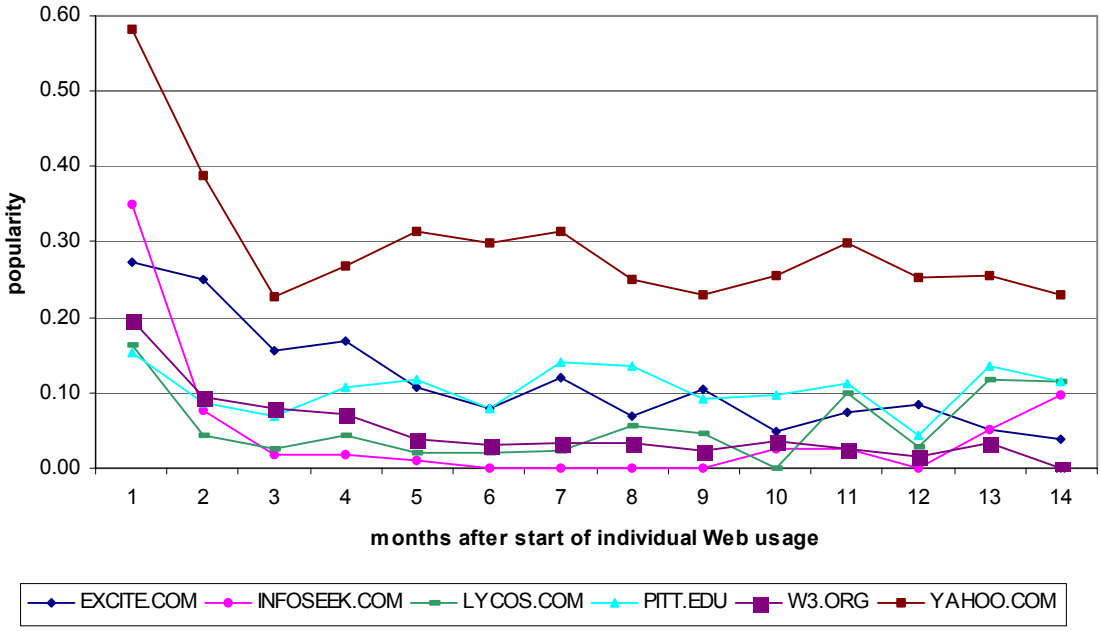


Figure 22: popularity over time

The popularity of another search engine, infoseek.com, develops differently over time. This Web site has a high popularity in the beginning, which is also reflected by a high overall popularity in Tab. 22, but fails to attract users again, which is also reflected by a low stickiness in Tab. 24. The popularity of this site decreases dramatically as shown in Figure 22.

Keep in mind that this relation (derived from a closed sample of HomeNet users) does not necessarily hold in the real world where there is no fixed set of users but a radical increase of the number of Web users. Even if individual users do not come back to a site, there are often enough new users to keep the site's popularity on a high level. However, the relation between stickiness and popularity gives us insights into the success or failure of some Web sites with low stickiness when the growth of the number of new Internet users slows down or even stops. As reported in [NUA], the number of new users accessing the Web is already saturationed.

5.5 The Popularity-Stickiness Map

Figure 23 displays the different values of popularity and stickiness on a popularity-stickiness map. Observe that this popularity-stickiness map displays only the most popular Web sites. Stickiness is depicted on the horizontal axis; popularity is depicted on the

vertical axis in a log scale. Because all the Web sites belong to the group of Web sites with the highest popularity, they are located in the upper half of Figure 23. If one chooses to display not only the popular but also all the other Web sites in the sample, this figure would look different. Given a power law of distribution of Web site visits as described in detail in [Adamic01], the figure would show a non-random distribution of Web sites across all possible values for popularity and stickiness with a cluster of Web sites in the lower left corner.

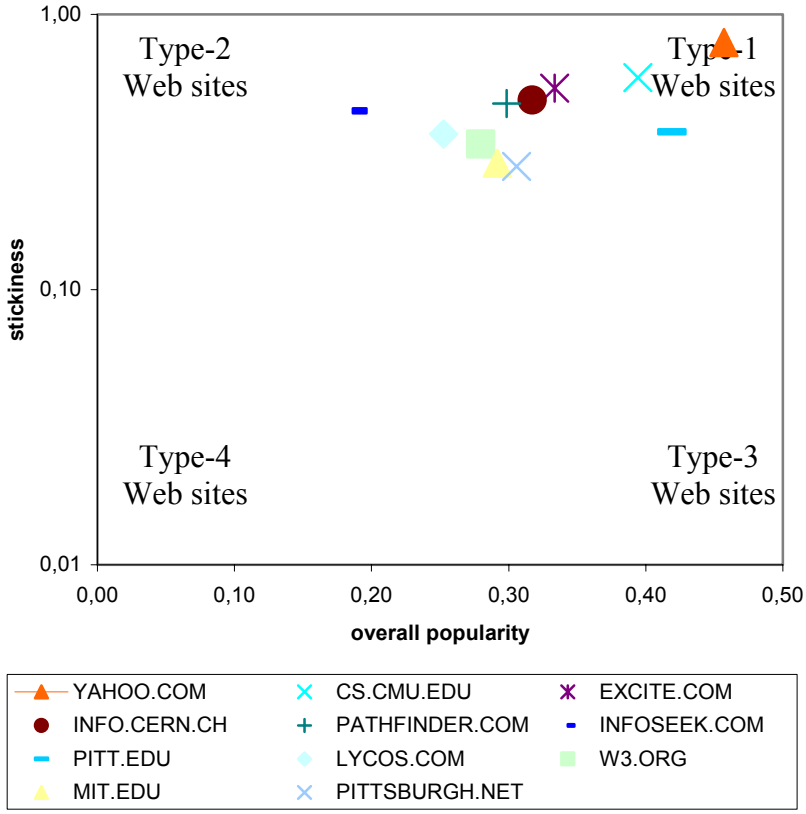


Figure 23: popularity-stickiness map of the more popular Web pages in the HomeNet sample

The popularity-stickiness map is divided into four areas with high or low values for stickiness and popularity respectively. The site that dominates all the other Web sites in Figure 23 is ‘yahoo.com’. Because both, popularity and stickiness are high, this Web site is located in an area of the upper right corner of the map and belongs therefore to the group labeled ’Type-1 Web sites’.

A direct comparison of ‘yahoo.com’ with other popular Web sites reveals that all other sites have both, lower (but still high) popularity and lower stickiness. The Web site that comes closest to a position in the upper left corner is ‘infoseek.com’. Web sites in this area of the map are labeled ‘Type-2 Web sites’. They succeed in attracting a lot of users but fail to make the same users come back to the site. Note, that the less sticky a site is, the more likely it is to become a ‘Type-3 Web site’ or even a ‘Type 4 Web site’ in the future when the growth of the Internet slows down and the Web site operators do not act to increase the Web site’s stickiness.

Some sites that have a much lower popularity can still be as sticky as ‘yahoo.com’, although for a limited number of users only. This group of Web sites are labeled ‘Type-3 Web sites’ because they address only a small subset of the population but achieve high stickiness among its users. In general, candidates for this category are Web sites that focus on a subgroup or niche of Web users and address the specific needs of these users, thereby providing high utility and achieving high stickiness.

‘Infoseek.com’ dominates sites which have the same small values for stickiness but smaller values for popularity, therefore positioning these sites closer to the even less desirable position in the lower left corner, an area in which ’Type-4 Web sites’ would be positioned. However, since Figure 23 displays only popular Web sites, none of the sites in this figure actually belongs to this group of Web sites. Candidates for this category are less popular sites with a low stickiness.

5.6 Dynamics of Web Site Popularity and Stickiness

The most desirable position for Web sites in Figure 23 is a position in the area of ‘type-1 sites’ because it permits using a variety of revenue models. However, given the results in section 5.4 and chapter 3 regarding saturation in Web usage and a high degree of disloyalty to Web sites, it is impossible for every Web site to reach this position. There will be a rather fierce competition of Web sites for the desirable positions in Figure 23. Given a degree of churn greater than zero, limited capacity of users turns the competition for desirable positions in Figure 23 into a zero sum game for competitors in the World Wide Web. In other words, whenever some Web sites improve their position in the popularity-stickiness map, others deteriorate their position. However, interesting niche strategies, such as ‘type-3-sites’ and - at least as long as the steady growth of the number of users in the Web continues - the ‘type-2-site’ strategy, can be the basis for revenue.

Figure 24 depicts possible paths of development for Web sites in the Internet. Path 6 and path 2 show how to reach a better position by increasing popularity. With regard to the results in section 5.2, this increases the probability of actually getting into the set of Web sites for given users at given points in time. Apart from word-of-mouth marketing, this can be done by pursuing customer acquisition strategies such as advertising, viral marketing, or referrals. One famous example for a Web site that transitioned from a ‘type-3 site’ to a ‘type-1 site’ is ‘hotmail.com’. Hotmail offered free e-mail for Web users. In the beginning, it had a limited but loyal customer base. Hotmail used viral marketing by adding a line to every e-mail that said ‘get your e-mail at hotmail.com’, thereby attracting more and more users. Network effects play an important role here. This strategy turned out to be very successful. Hotmail moved on path 2 to the area of ‘type-1 sites’.

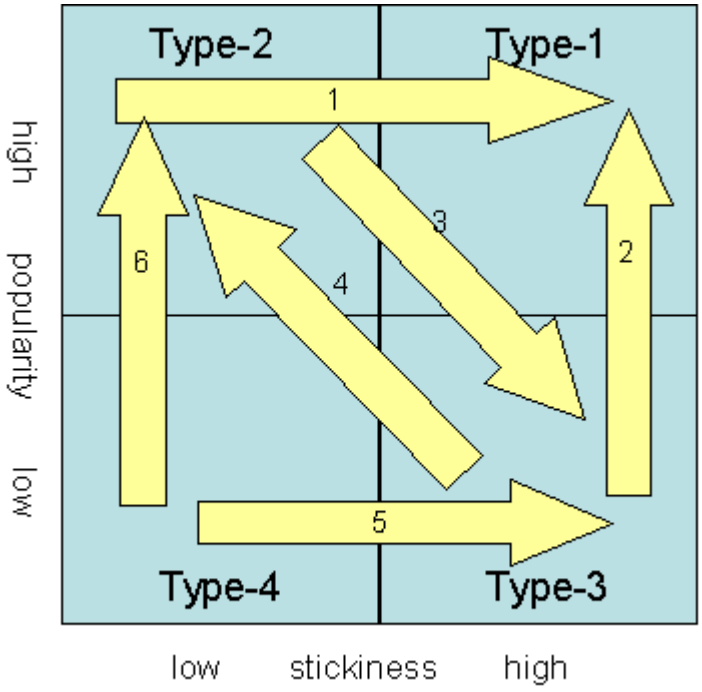


Figure 24: Dynamics of stickiness and popularity of Web sites

Path 1 and path 5 illustrate increasing stickiness of a Web site. With regard to the results in previous sections, this illustrated the entry of Web sites into the league of limited numbers of long-lived Web sites that actually stay in the set of Web sites of a given user. This can be done by pursuing customer retention strategies, such as personalization, which increase utility of Web sites and raises switching costs, e.g. by adding customized recommender systems to Web pages. Such recommendation systems are discussed in

detail in [Spiekermann00] and [Görsch00]. Because most individuals are interested in only a few of the millions of Web sites [Adamic01], such recommender systems can provide users with ‘easy ways to personalize their information space so that it reflects their interests’, which are demanded by [Kraut96b].

Another customer retention strategy is to decrease transaction costs by providing high quality service, e.g. ‘1 click ordering’ at ‘www.amazon.com’. With respect to the quality of a Web site, recent work by Yoo and Donthu [Yoo01] develops measures of site quality and hospitality to users. In a non-Internet marketplace customer loyalty is based to a large extent on customer trust and perceived service quality [Fukuyama95, Heskett94, Reichheld90]. Reichheld et al. [Reichheld00] suggest that this applies also to the customers of online vendors. Thus, Gefen [Gefen02] addresses the question whether the traditional market practice of maintaining customer loyalty by means of superior service quality and the trust that such service entails apply in electronic markets as well.

There is also the strategy of ‘community building’, which increases the utility to the users and builds barriers because of the difficulty to replicate community content. Famous examples for community content are, e.g., book reviews on ‘amazon.com’. Preset start pages of browser software are good examples for Web sites that followed path 1. In the early days of an increasing publicly available Internet, preset start pages of browsers such as ‘netscape.com’ for Netscape Navigator® and Microsoft Internet Exlorer® were candidates for the categories of ‘type-2 sites’. Because the start page was build into the software, every Internet user had to visit this page at least once until he changed the start page manually to a more appealing site such as ‘yahoo.com’. Browser companies reacted to this situation by increasing the stickiness of their Web sites. They added search engines first and transformed their sites to ‘portals’ with a variety of features later. Now users have less reason to change the start page of their Web browsers, which ultimately leads to a higher stickiness of these sites.

In the long run, ‘type-2 sites’ are in a difficult position because a substantial share of their users visits the site once but do not stick to it. Many of the ‘type-2-site’ users may be very new Web users. However, if the growth of the Internet in terms of number of users slows down, these Web sites are likely to change their position on the popularity-stickiness map. If following path 1 is not an option, these sites may consider pursuing a niche strategy to become a ‘type-3 site’. By focusing on specific user needs of a subgroup of users only, they might achieve a higher stickiness and follow path 3.

On the other hand, Web sites that already belong to the group of ‘type-3 sites’ may fail in their attempt to follow path 2 and follow path 4 instead, which puts them into the less

desirable position of ‘type-2 sites’. The common mistake they make is trying to reach a higher number of users by broadening their content or service. However, users who liked the site before the change because of its focus on certain content or service might find themselves in a new environment that does not necessarily meet their requirements anymore. These users might find it difficult to handle the broadened content or service and leave the site, e.g., because of increased search costs on this Web site. Ultimately, the stickiness of these Web sites will decrease. In this regard, personalization and customization techniques for Web sites that want to follow path 2 are in issues of utmost importance.

5.7 Conclusions Future Work

5.7.1 Major Results

This chapter is based on the findings from chapters 3 and 4 that emphasize that research on the dynamics of usage has to incorporate an analysis of churn in the Web. We address the question whether the groups identified in chapters 3 and 4 also differ in loyalty to Web sites and answer the question whether users converge over time to a set of ‘favorite’ Web sites. We develop key measures such as churn, popularity of Web sites and Web site stickiness. The major results are as follows:

- Users show little loyalty to Web sites. There is considerable churn in Web sites visited across subgroups of users.
- The degree of churn is a constant over time across all groups of users.

This is a surprising and interesting result and needs to be replicated in larger samples such as the Media Metrix panel usage data (see [Montgomery02]). It is important to identify Web sites that are both able to acquire and retain customers (popularity and stickiness) and to identify characteristics that contribute to the features. Given the fact that users reach saturation (as reported in the previous chapters) and show a constant high degree of churn over time, it seems relatively easy to get into a user’s set of Web sites. However, high churn also means that it is rather difficult to stay there. Therefore, we have analyzed which Web sites have the ability to get into this set of domains (popularity) and which have the ability to stay there (stickiness), and we have displayed Web sites in a ‘popularity-stickiness map’, which is divided into four areas. In this regard, the key results are as follows:

- Web sites differ substantially in popularity and stickiness. In particular, only a few Web sites are popular with most users.

- Among the popular Web sites, yahoo.com dominates the other sites with respect to popularity and stickiness.

Chapter 6 deals specifically with individual portal utilization at yahoo.com.

5.7.2 Implications for Electronic Commerce

One implication of the results, which indicated considerable churn across subgroups of residential users, is the opportunity to attract/acquire new customers (of course, retaining these customers is the difficult problem). We analyzed the ability of Web sites in the HomeNet sample to acquire and retain customers. Since users have a fixed number of sites that they are willing to visit in any give time period, we model both the popularity of a Web site (its acquisition ability) and its stickiness (its ability to retain customers).

The most desirable position for Web sites in Figure 23 is a position in the area of 'Type-1 Web sites'. However, given the results in chapter 3, the findings from the previous chapters on saturation in Web usage, and a high degree of disloyalty to Web sites, it is impossible for every Web site to reach this position. There will be a rather fierce competition of Web sites for the desirable positions in Figure 23. Given a degree of churn greater than zero, limited capacity of users turns competition into a zero sum game for competitors in the World Wide Web. In other words, whenever some Web sites improve their position in the popularity-stickiness map, others deteriorate their position.

Adamic et al. [Adamic01] report that the distribution of visitors per site follows a universal power law. The results from section 5.3 on popularity of Web sites conform to that. In the HomeNet data, only a few sites were broadly popular. The disproportionate distribution of user volume among sites is characteristic of winner-take-all markets [Frank95], in which the top few contenders capture significant market share [Adamic01].

In a non-Internet marketplace customer loyalty is based to a large extent on customer trust and perceived service quality [Fukuyama95, Heskett94, Reichheld90]. [Reichheld00] suggests that this applies also to the customers of online vendors. If that holds, limited loyalty in the online world could reveal a lack of service quality and trust in the Web.

The relation between stickiness and popularity gives us insights into the success or failure of some Web sites with low stickiness when the growth of the Internet in terms of number of new users slows down or even stops. For example, infoseek.com failed to reach its business goals in September 1998, which is already anticipated by the results on low stickiness of this site before 1998, as reported in section 5.4. Eventually, infoseek.com

was subject to a major redesign in 1998. At that time, personalized services were added to the site.[30]

It is crucial to understand that the conclusions from the findings on stickiness and popularity of Web sites are not to increase traffic at a given Web site at all cost. Traffic at a Web site tell us little about its success. Using the two measures of popularity and stickiness combined provides a much more accurate view on the Web site's success than just counting page views.

5.7.3 Future Work

We believe that future work is needed which takes into account human context when looking at churn in terms of both why users choose particular sites and what constitutes disloyalty. For example, infrequent use of the same site does not necessarily constitute disloyalty. In this regard, the measure of churn used in this work is a simple one in the sense that it does not incorporate the type of Web site. For example, some types of Web sites, such as vacation sites, are by nature visited infrequently. However, users might still be loyal to these sites.

Relatedly, users may visit Web sites that are functionally related, e.g. vacation sites. If in fact this is the case, there is a need to develop methods for modeling churn, which take into account the possibility that Web sites maybe complements and substitutes to one other.

Future research should also try to answer the question if there are particular Web site attributes that increase or decrease Web user loyalty.

[30] Personalization based on Web usage data has been discussed in [Mobasher01].

6 Portal Utilization

Based on the findings in the previous chapters, this chapter advances the research on Web usage by specifically addressing the issue of Web portal utilization. The previous chapter 5: 'Web User Loyalty and Web Site Stickiness' reveals that the Web portal yahoo.com is among the most popular sites in the HomeNet sample. In fact, portal sites are becoming increasingly popular in the World Wide Web. In this chapter, we measure Web portal utilization of individuals and develop demographic profiles of user groups with different portal utilization levels.

6.1 Introduction and Motivation

Many Web users set their browser starting page to Web portals such as yahoo.com or excite.com [Wiggins01]. Many of these portals began their existence as robotic Web indexers. Since that time, most popular search engines morphed into portals. For example, they added human-edited content, such as news, and additional services, such as yellow pages. Sites that used to offer only search capabilities made the transition to Web portals by incorporating such additional services. However, even successful businesses such as yahoo.com currently face difficulties in maintaining their revenue stream [Hansell01a]. Understanding the customer of portal sites may help to secure profitability and enhance the design of the site. Therefore, the objective of this analysis is to better understand the portal utilization rates of individual users. Specifically, this chapter reports the results of an analysis of the HomeNet data on individual Web portal usage. This data was collected between 1995 and 1997, the very period of time in which many navigational directories and search engines became Web portals. We measure Web portal utilization of individuals and develop demographic profiles of user groups with different portal utilization levels.

Because Web portals need to know who their customers are, this analysis addresses the question what demographic characteristics distinguish Web users who actually make use of the additional features of portal sites such as yahoo.com on the one hand and Web users who use the search capabilities only on the other hand. We also compare the demographic profiles of user groups with different portal utilization levels with demographic profiles of groups with different overall Web utilization levels to test if a digital divide that exists with respect to overall Web usage (see chapter 3: 'Saturation of Lay Web Usage' on pp. 40 ff.) also exists with respect to Web portal utilization.

We focus on portal utilization on a specific portal site: yahoo.com. Yahoo.com was one of the first online navigational guides to the Web. Over time, it developed strong brand recognition and grew tremendously. For example, yahoo.com has seen exponential increases in page views per day over the period of observation. The analysis of Web usage of individuals in the HomeNet project conducted in the previous chapter confirms that yahoo.com is by far the portal site with the highest popularity and brand recognition (see also Tab. 22: Most popular sites in the HomeNet sample on page 87). According to self-reported usage data from yahoo.com [Yahoo99a], the average number of page views per day grew from 9 million in June 1996 to 65 million in December 1997. Chapter 5 also revealed that Web users tend to visit yahoo.com repeatedly. All this turns yahoo.com into a welcome business partner for companies, which want to place banner advertisements. The number of advertisers on yahoo.com grew from 230 in June 1996 to 1700 in December 1997 [Yahoo99a]. However, these overall numbers do not tell us how successful the additional services on yahoo.com actually are. Because the utilization rates of the additional services among users clearly influence how suited these services for banner advertisements are, the aim of this analysis is to measure the utilization rates of users with respect to these additional features at yahoo.com. Perhaps, additional services available to users are not easily realizable, which would lead to a low utilization rate for these services.

The structure of the chapter is as follows: Section 6.2 discusses our measure of portal utilization: The count of services used by individuals. Section 6.3 provides the results of this study. Section 6.4 discusses the results, provides important implications for electronic commerce and related public policy, and concludes the chapter by discussing future work issues.

6.2 Data Extraction and Method

Using data from the HomeNet project (see chapter 2), we measured Web portal utilization by the number of distinct services visited at yahoo.com by given users. Yahoo.com offers a variety of services, which include:

- search (search.yahoo.com),
- headlines (headlines.yahoo.com),
- finance/business (biz.yahoo.com or finance.yahoo.com),
- maps/driving directions (maps.yahoo.com),

- yellow pages (yp.yahoo.com),
- chat (chat.yahoo.com),
- local services (la.yahoo.com or ny.yahoo.com),
- and many others.

In this respect, by measuring how many services people actually visited we measured how broad a user's interest in a portal site was rather than measuring the intensity of usage at this site. Yahoo.com's Web server design allows for identifying the services visited by a given user simply by extracting the URL from the log file. For example, maps and driving directions can be found at maps.yahoo.com, whereas yellow pages can be found at yp.yahoo.com. We measured the number of distinct portal service URLs visited by given individuals in the period of observation. For example, users who visited search.yahoo.com and maps.yahoo.com utilized two services. To avoid skewed results, a service is not counted if it was accessed first during an Internet session, which indicates that the user defined this service as his start page. However, such a service is counted if the same user accessed it later on during a session.

We used regression analysis to identify the characteristics of users that predict portal utilization differences. The dependent variable in this analysis is: the count of Web services visited at yahoo.com by individuals. The predictor variables are demographic variables provided by the HomeNet data, which were measured on questionnaires.

6.3 Results

As a first step, we examined how many people in the sample were actually making use of the additional services at yahoo.com. Only 24% of the users visited more than one service at yahoo.com. Only 13% of the users visited more than two services. We wanted to test whether the transition of navigational directories and search engines to portal sites automatically led to heavy use of these additional services. Judging from the results, this does not seem to be the case.

In order to identify characteristics that distinguished individuals in these groups, we examined the demographic profiles of the groups with portal utilization levels of 0, 1, 2, and greater than 2 services used. Tab. 25 shows the demographic differences across the various groups. The summary statistics reveal that individuals in the groups that used portals heavily (portal utilization >=3) tend to be older and male. Conversely, groups that

made little use of portals (portal utilization=0) were disproportionately comprised of adult females and minorities.

Tab. 25: Overview of characteristics of users in the various groups

	All users	0 services	1 service	2 services	>=3 services
Percentage	100%	20.14%	54.68%	11.51%	12.95%
Female	51.80%	65.52%	48.68%	62.50%	33.33%
Minority	26.62%	41.38%	19.74%	31.25%	27.78%
Position:					
Mom	25.90%	34.48%	25.00%	18.75%	30.77%
Dad	20.14%	3.45%	23.68%	18.75%	15.38%
Daughter	20.14%	20.69%	21.05%	31.25%	7.69%
Son	17.27%	13.79%	19.74%	12.50%	23.08%
Other	16.55%	27.59%	10.53%	18.75%	23.08%
Avg. age	32.63y	32.93y	31.28y	31.36y	38.78y

Figure 25 displays the distribution of the number of portal services used by individuals of different age in a scatter diagram. Clearly, the majority of Web users make use of one service only. The number of users decreases as portal utilization increases. For example, 54.68% of the individuals in the project visited only one service whereas only 12.95% of the people visited more than two services. No individual visited more than 5 services.

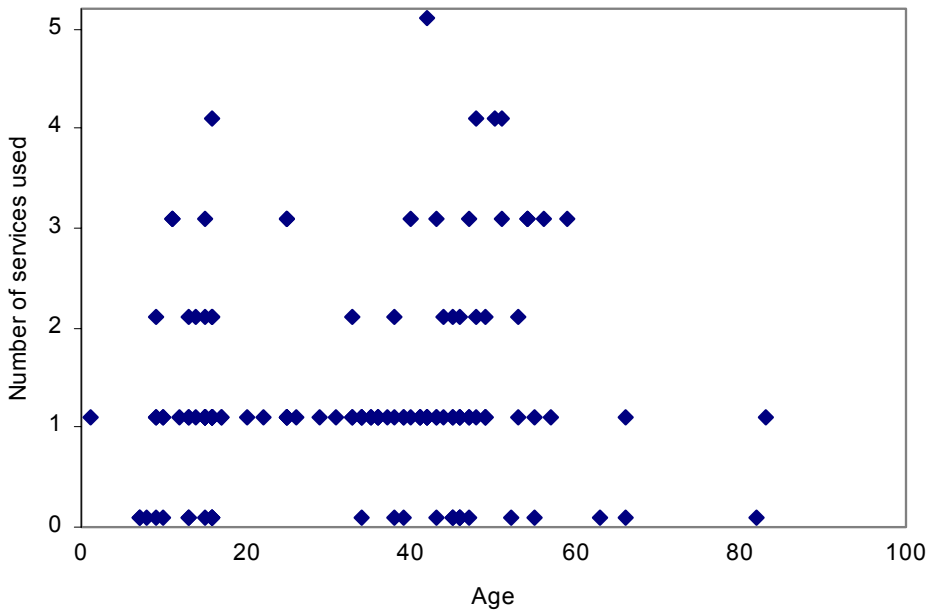


Figure 25: Distribution of portal utilization and age

Figure 26 - Figure 27 visualize these results. The findings are first discussed informally, followed by a formal regression analysis.

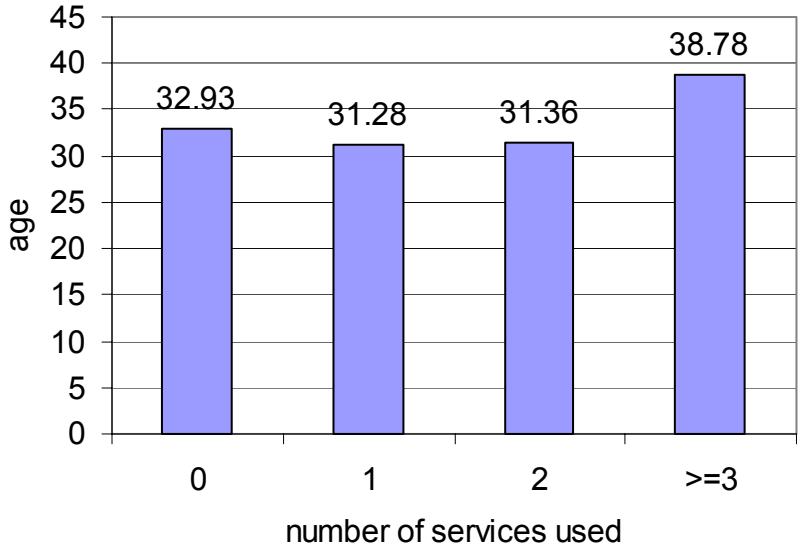


Figure 26: Average age of individuals in groups with different portal utilizations

As shown in Figure 26, the average age of individuals in the groups with portal utilization levels of 0, 1, and 2 services used stays almost constant. However, heavy users with a portal utilization greater than 2 services used tend to be significantly older. Judging from Figure 26, age has a positive impact on heavy portal utilization.

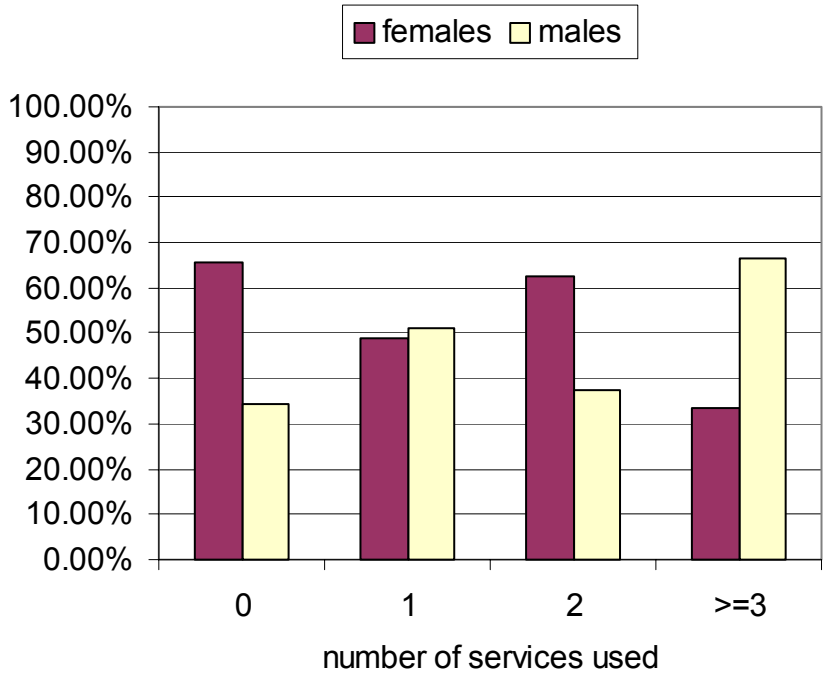


Figure 27: Percentages of males and females in the groups with different portal utilization levels

Figure 27 reports how gender determines portal utilization. It seems to be the case that being male has a slight positive impact on portal utilization. The following figure reveals if race determines Web portal utilization.

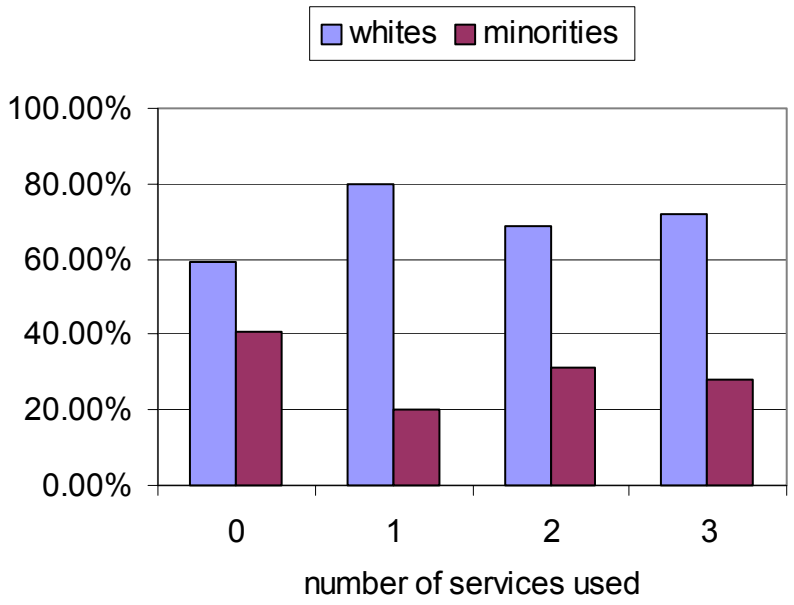


Figure 28: Percentages of whites and minorities in the groups with different portal utilization levels

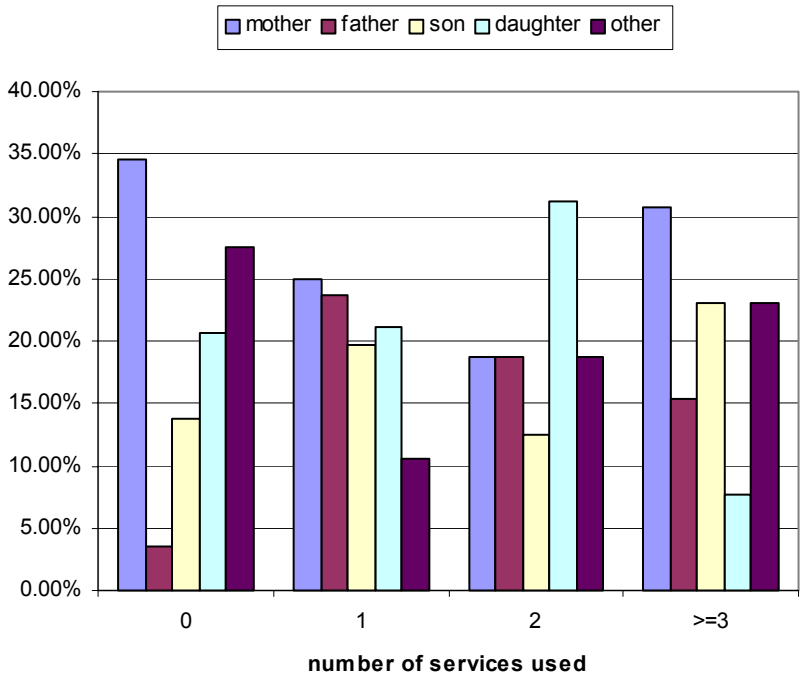


Figure 29: Roles in the family of individuals in the groups with different portal utilization levels

Figure 29 depicts the percentages of people in the project who play different roles in the family and their portal utilizations. Apart from the finding that non-users tend to belong to

the group of ‘mothers’, no significant differences for usage levels of 1, 2, and 3 services used can be identified.

So far, it seems that only some demographic characteristics such as age and gender determine portal utilization. To test for the robustness and significance of these results, we conducted a regression analysis to evaluate how race, age, and gender, influence Web portal utilization. Specifically, we applied a poisson model for which we chose the dependent variable ‘number of portal services used’.

$$SERVICECOUNT = \alpha+\beta_1 WHITE+\beta_2 ELDER+\beta_3 FEMALE$$

Using a model whose underlying statistical assumption is a poisson distribution is very common when the dependent variable is a count. The following binary variables are the parameters of the model:

- ‘white’, the race of the individual,
- ‘elder’, the age group of the individual; ‘elder’ is ‘true’ if the indivual’s age is greater or equal 40 year,
- ‘female’, the gender of the individual.

The regression aimed to test which demographic characteristics determine portal utilization. Tab. 26 reports the results of this analysis.

Tab. 26: POISSON estimates

```
Poisson regression                Number of obs   =        139
LR chi2(3)      =      6.49
Prob > chi2     =    0.0901
Log likelihood = -184.80    Pseudo R2       =    0.0173

------------------------------------------------------------
portutil |  Coef.  St.Err.    z   P>|z|  [90%Conf.Int]
-------------+----------------------------------------------
female |  -.304   .1586  -1.92  0.055  -.5652  -.0434
white |   .080  .18283   0.44  0.659  -.2200   .3814
elder |   .292    .168   1.74  0.082   .0157   .5687
_cons |   .193  .19005   1.02  0.309  -.1193   .5058
------------------------------------------------------------
```

Judging from the results in Tab. 26, it seems that race has no significant impact on portal utilization in the World Wide Web! This is an important and surprising result given the fact that the ethnic background does have a significant impact on overall Web utilization as

reported in previous chapters. It seems that portal sites speak more to minorities and women than other sites in the Web.

Gender and age determine usage to some extent. Being female seems to decrease the probability of a given user to have a high portal utilization, whereas being older seems to increase this probability.

We also tested the robustness of the finding that role in the family has no significant impact on portal utilization with a regression analysis. This analysis did not reveal such significant estimates.

As a next step, we were interested in the question how overall Web utilization affects specific portal utilization. We have identified four groups with distinct trajectories of overall Web usage over time in the HomeNet sample in chapter 3. According to the overall Web utilization rates identified in that chapter, which were measured in number of distinctive Web sites visited per month by given users, these groups were labeled 'light users', 'moderate users', 'heavy users', and 'very heavy users' respectively. It was reasonable to expect that the very fact that portals aggregate content leads to a low overall Web utilization rate if measured in the number of distinctive Web sites accessed because portal users find all they need on one portal site. Therefore, we tested if there is in fact a negative correlation between overall Web utilization and portal utilization. This analysis actually revealed a moderate positive correlation with a correlation coefficient of 0.3. Judging from this result, it seems that portal usage does not increases as overall Web utilization decreases.

6.4 Conclusions and Future Work

6.4.1 Major Results

This chapter extends the work of prior studies by presenting the results of a long-term study with residential subjects. It specifically addresses the issue of Web portal utilization at a specific portal, yahoo.com. It thereby answers the question if Web portal users are different from average Web users.

The major results are as follows:

- Only 23% of the users visit more than one service at yahoo.com, only 13% visit more than two services.
- The ethnic background of individuals has no significant impact on portal utilization in the World Wide Web.

- Males and elder people are more likely to show high degrees of portal utilization.

The results have implications for electronic commerce and public policy that are discussed in the subsequent sections.

6.4.2 Implications for Electronic Commerce

The analysis presented in this chapter reveals that only 23% of the users visited more than one service on yahoo.com over the period of observation. Only 13% of the users visited more than two services. Note that Web portals invested a lot of money in these rarely used services. One reason for the low utilization rates may be a marketing problem of yahoo.com. It seems that services available to users were either not very easily realizable to users or just not attractive enough to users over the period of observation.

We also tested whether a link existed between overall Web utilization and portal utilization. Maybe people with low overall Web utilization show a high portal utilization as well. On the other hand, it is reasonable to expect that portal users have a lower overall Web utilization if measured in number of distinct sites accessed because they satisfy their variety of needs on a single portal site instead of going to a variety of sites. The results show that this is not the case; overall Web utilization had a slight positive impact on portal utilization.

Apart from identifying groups of portal utilization, we discovered distinctive demographic factors that distinguish these groups. The results show that gender and age determine Web portal utilization. However, there is no such thing like a ‘golden customer’ for portal sites with respect to the characteristics race and role in the family. The impact of age and gender is statistically significant, but the estimates calculated in the regression analysis show that this impact is minor.

6.4.3 Implications for Public Policy

Apart from the fact that the identification of demographic factors that characterize the groups of portal utilization provides valuable insights from a marketing perspective, the results reported in this study also have importance from a public policy perspective. In particular, they speak to the digital divide debate, which is discussed in detail in chapter 7: ’The Digital Divide Exists’ on pp. 111 ff.

It was reasonable to examine if a possible race or gender gap exists with respect to portal usage, and to find out if a possible digital divide exists with respect to portal utilization to

the same extent as it exists with respect to overall Web usage in the period of observation. Judging from the results of this analysis, a gender gap exists to some extent. For example, the percentage of females is 51% in the overall sample, 66% for non-users, and 33% for very heavy users with utilization levels greater than 3 services. This confirms the results of other studies, which identified a gender gap with respect to overall Web utilization. However, even though a gender gap still existed in the HomeNet sample with respect to Web portal utilization, this gap is by far not as big as the gender gap with respect to overall Web usage. There was no race gap with regard to portal utilization.

According to studies on the digital divide conducted in previous chapters, such heavy and very heavy Web users tend to be younger, male, and white. Even though there is a positive correlation between overall Web usage and portal utilization, this analysis does not confirm that heavy portal users share the same demographic characteristics with people with high overall Web utilization. There is less of a digital divide with respect to portal usage than with respect to overall Web usage. Thus, it seems that portal sites speak more to minorities and women than other sites in the Web. Also, the finding that the role in the family has no significant impact on portal usage extends the findings on overall Web usage such as [Kraut96a], which show that teenagers, particularly boys, use the Internet more intensely than their parents. Apart from the slight impact of age and gender on usage, the demographic characteristics are surprisingly similar across the groups with different portal utilization. Moreover, in contrast to the results of studies on the digital divide, which show that heavy and very heavy users tend to be younger, age even has a positive impact on portal usage in this analysis.

6.4.4 Future Work

Future work seems necessary with respect to a variety of issues. It is important to keep in mind that the low portal utilization at yahoo.com may simply mean that users are going elsewhere. Therefore, utilizations of competing portals should be measured in future studies.

Future research with respect to the identification of the exact services used by individuals is desirable. So far we measure only the degree of portal utilization without saying anything about the identity of the services used. Relatedly, one might think of testing which additional services actually contribute to an increase in the popularity of the portal site. In particular, personalized services such as my.yahoo.com are supposed to make users more loyal to a site and increase popularity.

In particular, different demographic segments should be the focus of further analysis. For example, mothers seem to make up a large part of the heavy user group. It would be interesting to learn more about the usage pattern of this specific group, as it is commonly underrepresented in similar contexts.

Finally, it is important to measure portal utilization not only by number of services used but also by number of page views with respect to the specific services visited. Note that the measure used in this chapter, the number of services visited on yahoo.com, leaves out the intensity as in repeat visits. The number of page views and how it evolves over time has important implications for Web sites with advertisement models that rely on a high number of page impressions, such as yahoo.com. The number of page views also indicates how satisfied visitors are with a given service.

7 The Digital Divide Exists

7.1 Introduction

As we have seen in the previous chapters, many of the results presented in this dissertation have implications for public policy as it pertains to the digital divide. As the Internet has grown and become more widely used by government and organizations, concerns have been raised about this issue [NTIAa]. This chapter summarizes empirical research on the digital divide in the United States and Europe.

The digital divide was coined by former United States Assistant Secretary of Commerce for Telecommunications and Information, Larry Irving [Miller01]. It refers to the gap between those members of the society who can effectively use new information and communication tools, such as the Internet, and those who cannot. Those who cannot may be unable to benefit from these new information tools due to their lack of access or due to their inability to make full use of it. Researchers are nearly unanimous in acknowledging that some sort of divide exists at this point in time [NTIA02].

William M. Duley, former head of the United States department of commerce, highlights the importance of bridging the digital divide as follows:

“A country’s most important resource is its people. Companies are only as good as their workers. Highly skilled, well-educated workers create superior products. In a society that relies on the Internet to deliver information, it is important that everybody has access, which in turn will produce then a technology-literate work force.” [NTIA02]

In 2000, the U.S. Census Bureau reported that 281.4 million people lived in the United States [USCensus00].[31] People who are unable to participate in electronic commerce cannot benefit from the convenience and tax savings of online shopping in the United States. Businesses are unable to set up online sales operations and local authorities cannot offer services online. In general, unequal adoption of technology excludes many from reaping the fruits of the economy [DDN]. For example, Figure 30 depicts the favorite online activities of Internet users across age groups. Clearly, not having access to services such as ‘employment search’ in the Internet is a crucial disadvantage to individuals who do not use the new Information technology.

---

[31] Among them 138.1 million males, 143.4 million females, 211.5 million Whites, 34.7 million Blacks, 35.3 million Hispanics, 72.3 million people under age 18 and 35.0 million people over age 65.

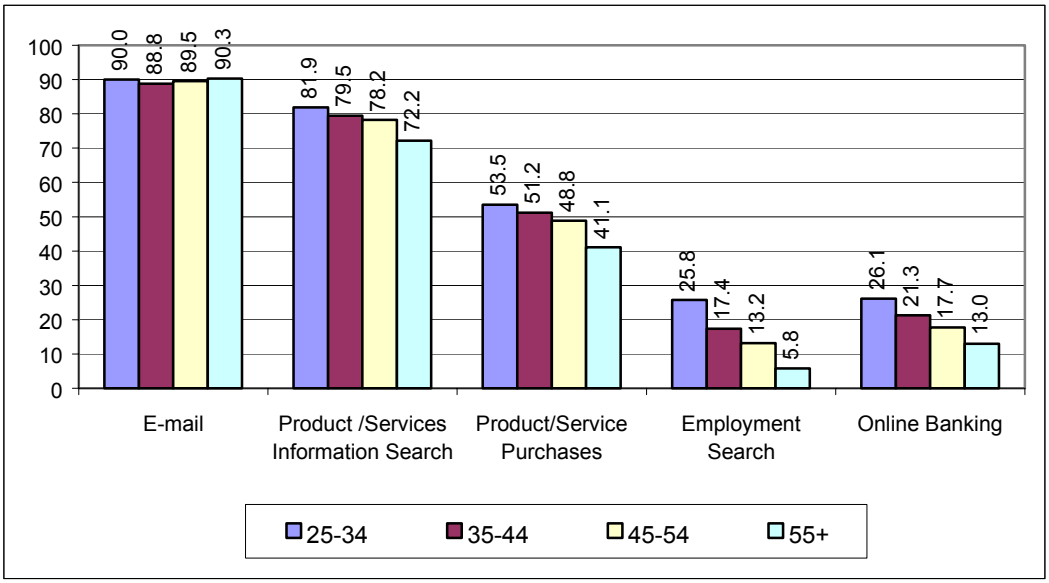


Figure 30: Selected online activity by age (2001) [NTIA02]

Therefore, as far back as 1995, the United States government issued first reports on the digital divide, such as 'Falling through the Net: A survey of the 'Have-Nots' in Rural and Urban America' [NTIA95]. These and subsequent reports [NTIAa, NTIA99b, NTIA02] are a first step to understand, measure, and explain how the information revolution is affecting people. They provide the factual foundation for key policy initiatives to promote access to the Internet for all groups of people. Such initiatives provide, e.g. Internet access at schools. By 2000, the United States government invested 5.5 billion US dollar per year in the information infrastructure [Perillieux00].

Based upon these reports, research studies such as [Hoffman98a] have carefully examined the policy implications of demographic patterns of Web usage. They provide evidence that the 'digital divide' between certain demographic groups and regions continues to persist and in many cases is even widening.

The digital divide can be studied from a global or national perspective. Because the data set used for this dissertation was collected from Americans from the Pittsburgh area, section 7.2 focuses on describing the issue of the digital divide from an American perspective. The digital divide in a European and German context is discussed in section 7.3. Section 7.4 briefly puts the issue of the digital divide in a global context. Finally, section 7.5 summarizes key implications drawn from the current research findings.

7.2 The Digital divide in the United States

The spread of new technologies, such as the Internet, can be described by a variety of metrics. For example, the percentage of individuals and United States households connected (see Figure 31 and Figure 32).

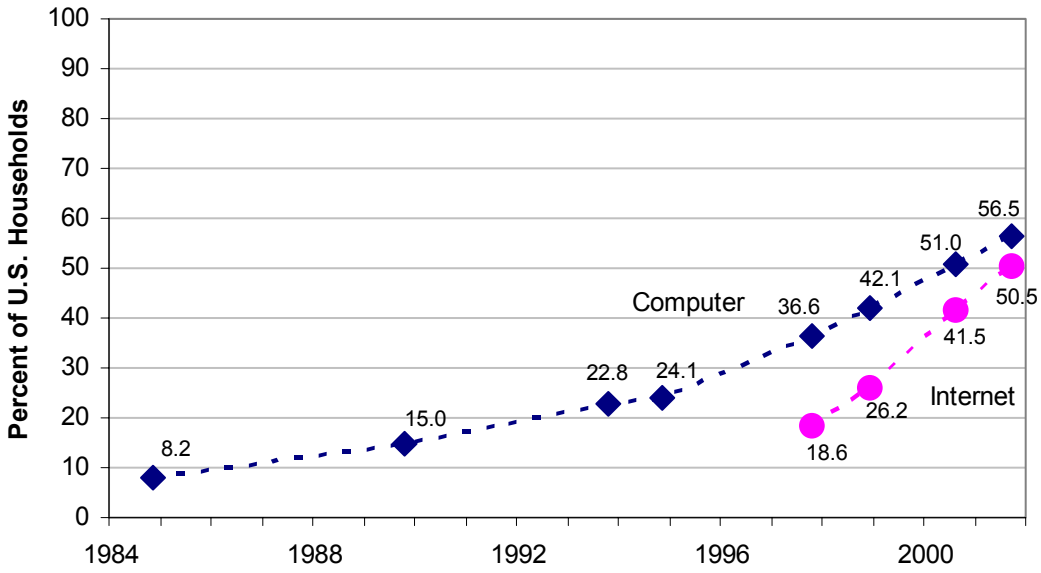


Figure 31: Percentage of United States households with a computer and Internet connections [NTIA02]

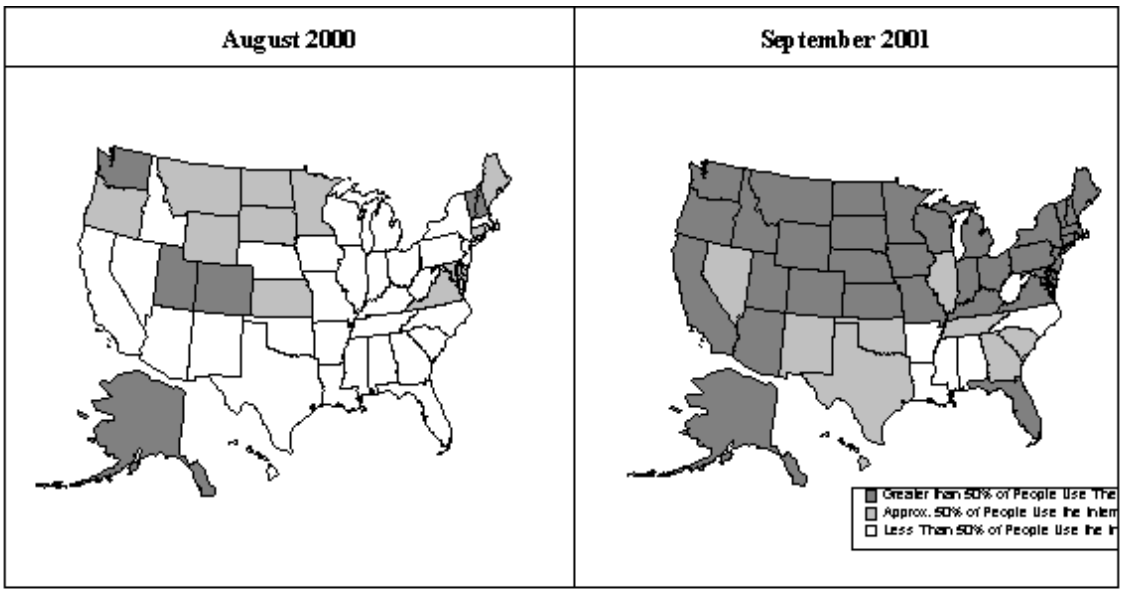


Figure 32: The rapid increase in Internet use in the United States across states [NTIA02]

On average, information technology penetration is rising. In particular, significant changes have occurred for personal computer ownership and Internet access. In 2002, 51% of all United States' homes had a computer and 41.5% of all United States homes had Internet access [NTIA02].

However, these numbers on computer ownership and Internet usage are averages that may conceal substantial heterogeneity. Internet connectivity of certain groups may be growing more rapidly. The chief concern with respect to information technology penetration is that groups that were already connected are now far more connected, while those with lower rates have increased less quickly, which results in a growing gap between information "haves" and "have nots" over time [NTIA99b]. Therefore, a deeper analysis is necessary with respect to specific characteristics that determine access and usage. With respect to information technology, penetration and usage levels differ substantially according to income, education level, race, geography, and other demographic characteristics [NTIA02]. Specifically, although the digital divide is closing in many different aspects, it is widening in other aspects. Numerous studies such as [Cyb98a] and [Abrams97a] predicted already in 1998 that while the gender gap in the United States will likely close over time, the race gap will prevail. This confirms our findings on race and gender gaps with respect to Web usage as presented in the previous chapters.

The following key characteristics that determine Internet access and usage are commonly acknowledged in the research literature [Cyb98a, NTIAa, NTIA99b, NTIA95, NTIA02, Hoffman98a, DDN]: race, gender, educational attainment, and age of individuals as well as the geographic region they live in. The following sections summarize the findings of studies that analyzed the impact of these characteristics on computer and Internet use.

7.2.1 Divide Based on Race/Origin

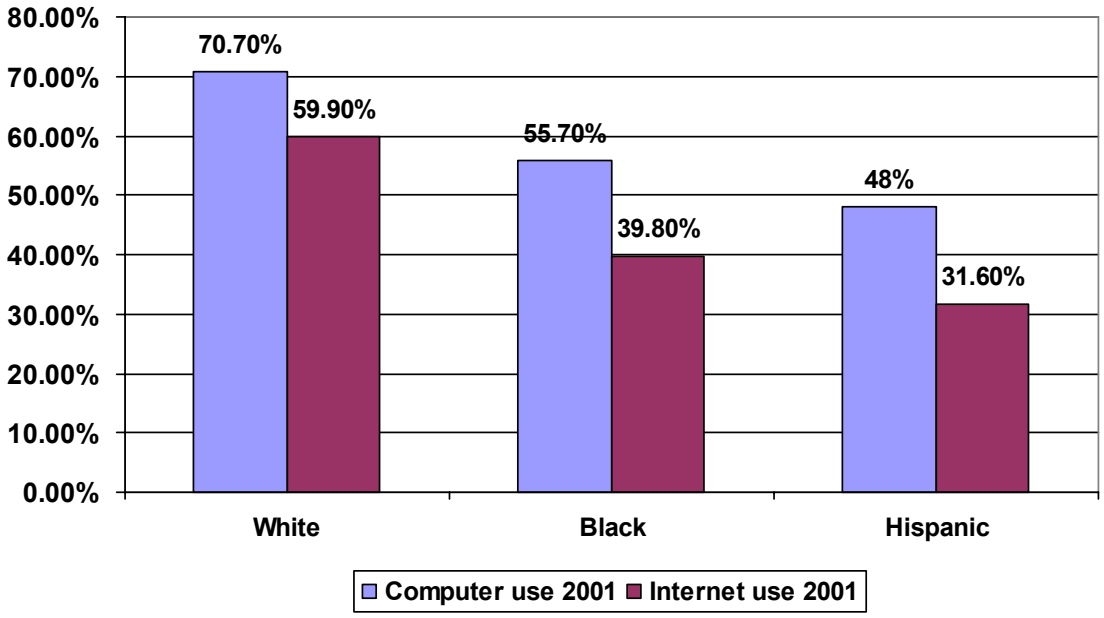


Figure 33: Internet use and computer use in the United States by race / Hispanic origin [NTIA02]

Already in 1998, Novak et al. [Novak98] reported that whites are significantly more likely than African-Americans to have a home computer in their household and to have access to a personal computer at work. Moreover, whites are significantly more likely to have ever used the Web at home.

In 2002, whites are still more likely to have access to the Internet from home than Blacks or Hispanics. White (59.9%) households continued to have Internet access at significantly higher levels than Black (39.8%) and Hispanic (31.6%) households. Also, home computers are less common in minority households (white 70.7%, black 55.7%, Hispanic 48%), even controlling for education and income [Anderson95, Rockman95, NTIA99b,

NTIA02] (See also Figure 34). There is still a yawning divide among different races and origins.

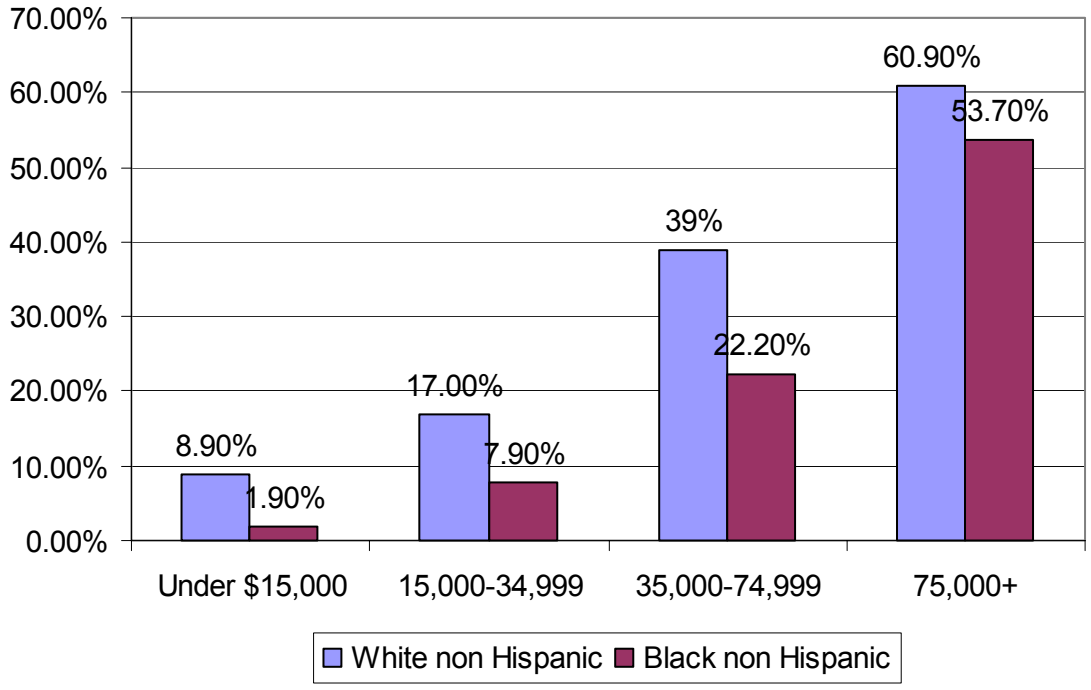


Figure 34: Internet use in the United States (race and income) [NTIA99b]

[Novak98] specifically addresses the questions how computer ownership and access (both are determined by race/origin) relate to Web usage. Individuals who own a home computer and have access to a computer at work are much more likely than any other group to have used the Web. In this regard, computer access seems to explain subsequent Web usage.

In this regard, the analysis in the previous chapters of this dissertation tests if race differences in the HomeNet sample determine Web usage. Many of the findings confirm the current research on the digital divide. However, new research findings can be presented that lead to new conclusions. For example, chapter 3 reports that race differences remain even if equal Internet and computer access is given to all people.

7.2.2 Divide Based on Gender

By 2002, gender differences explaining computer and Internet access have vanished. In 1997, males were more likely than females to be Internet users [NTIA02]. Figure 35 and Figure 36 report computer use and Internet use by gender.

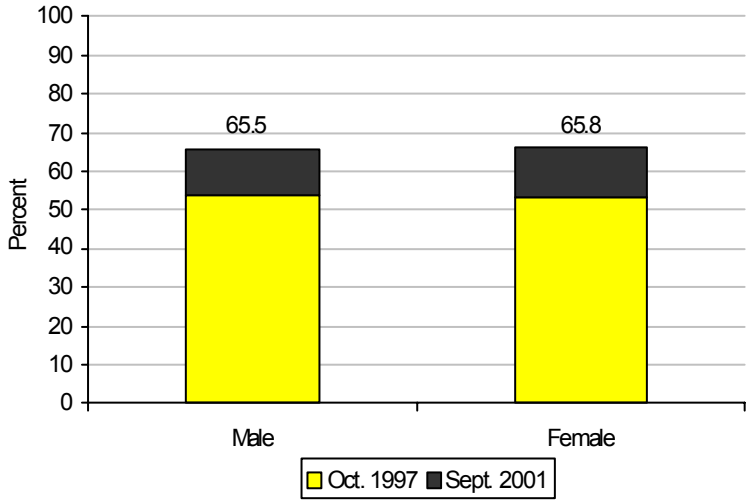


Figure 35: Computer use in the United States by gender [NTIA02]

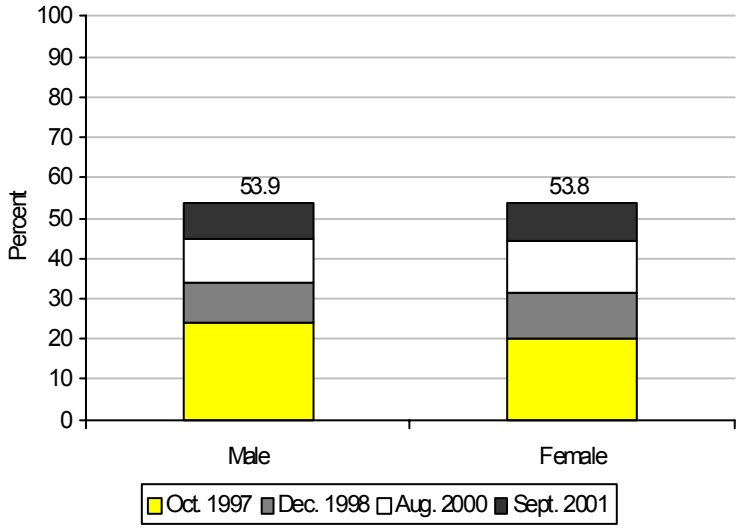


Figure 36: Internet use in the United States by gender [NTIA02]

Until recently, females often seemed not as attracted to computers as men did. Societal, familial, and educational attitudes worked against women in computing [Klawe95]. In this regard, the results reported in previous chapters which rely on data from 1995-1997 are

historical testimonies of these days, because they still report a gender gap that even prevailed after barriers to usage were reduced.

7.2.3 Divide Based on Income

Until recently, few families with lower incomes owned a computer (see also [Anderson95]). In 2002, 86.3% of households earning $75,000 and above per year had Internet access compared to 12.7% of households earning less than $15,000 per year. Figure 37 and Figure 38 depict computer use and Internet use by family income. Judging from these figures, the income gap is clearly not closing over time.

Already in 1999, one of the 'Falling Through The Net' reports issued by the United States government [NTIA99b] revealed that "Urban households with incomes of $75,000 and higher are more than *twenty* times more likely to have access to the Internet than rural households at the lowest income levels, and more than *nine times* as likely to have a computer at home".

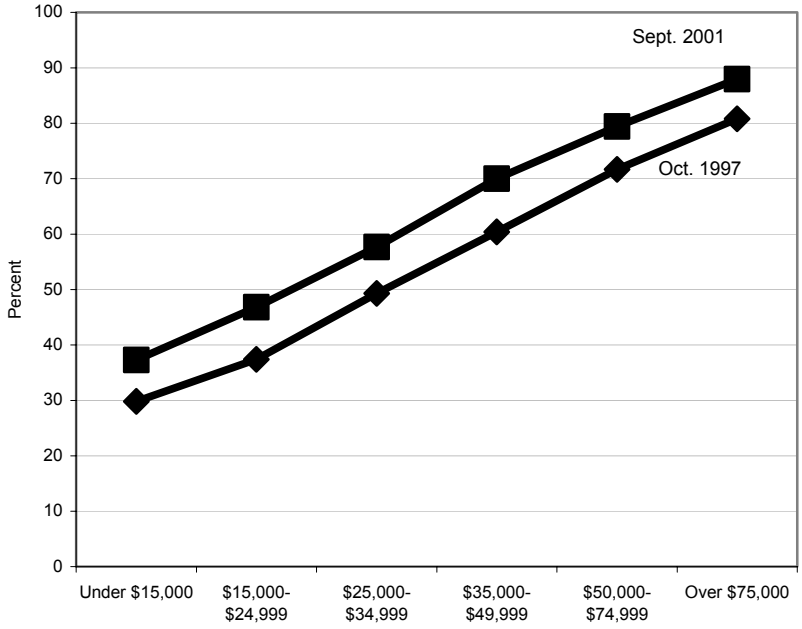


Figure 37: Computer use in the United States by family income [NTIA02]

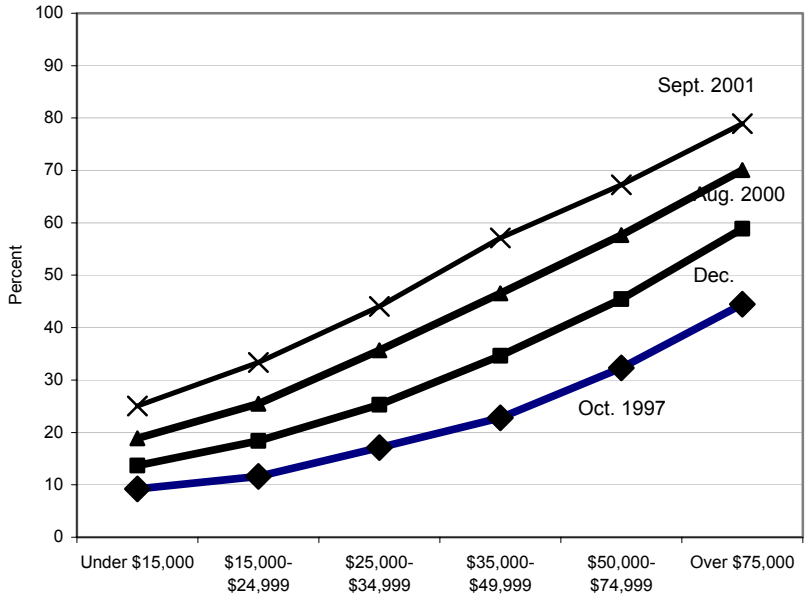


Figure 38: Internet use in the United States by family income [NTIA02]

Figure 39 reveals that income determines the awareness of Internet usage cost. As depicted in Figure 46, cost and the investment required to purchase a computer are still considerable barriers to Internet usage.

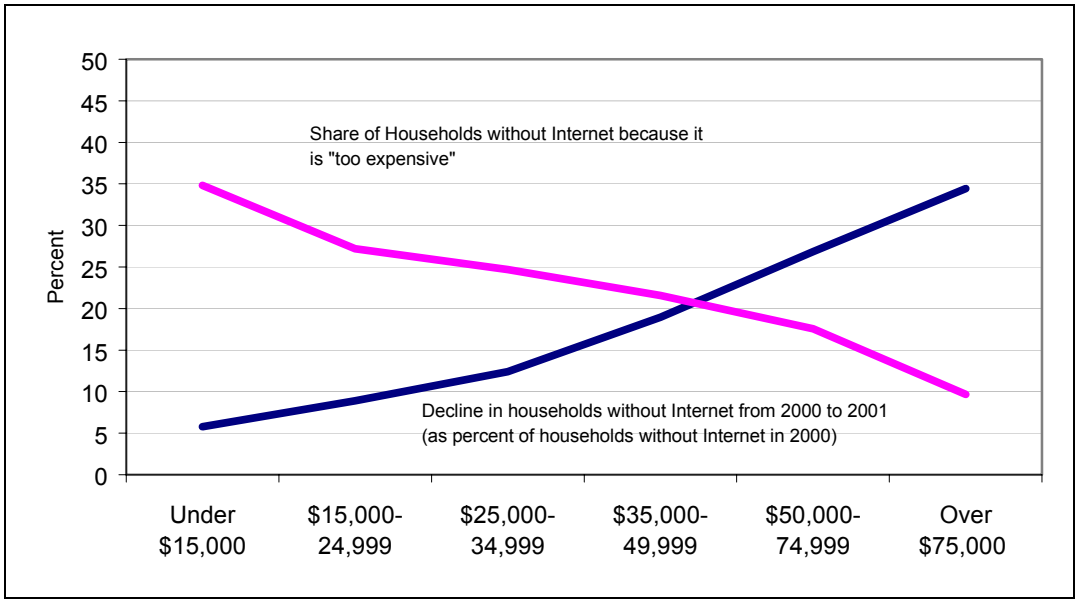


Figure 39: Adoption rate and Internet "too expensive" by income percent of United States households without Internet [NTIA02]

Notice that the hurdle of acquiring a new computer was removed in the HomeNet project. All users in the HomeNet sample received a computer for free. As reported in previous chapters, household income did not determine Internet usage of HomeNet participants.

7.2.4 Divide Based on Education Attainment

Judging from Figure 40 and Figure 41, there is an apparent effect of education on computer use and Internet use. Nearly 65% of college graduates have home Internet access; only 11.7% of households headed by persons with less than a high school education have Internet access.

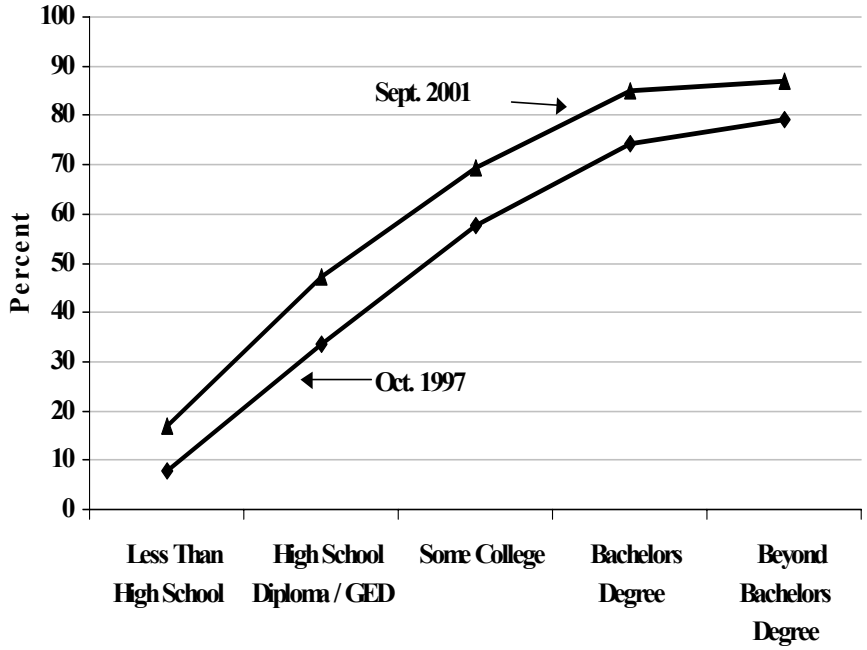


Figure 40: Computer use by education [NTIA02]

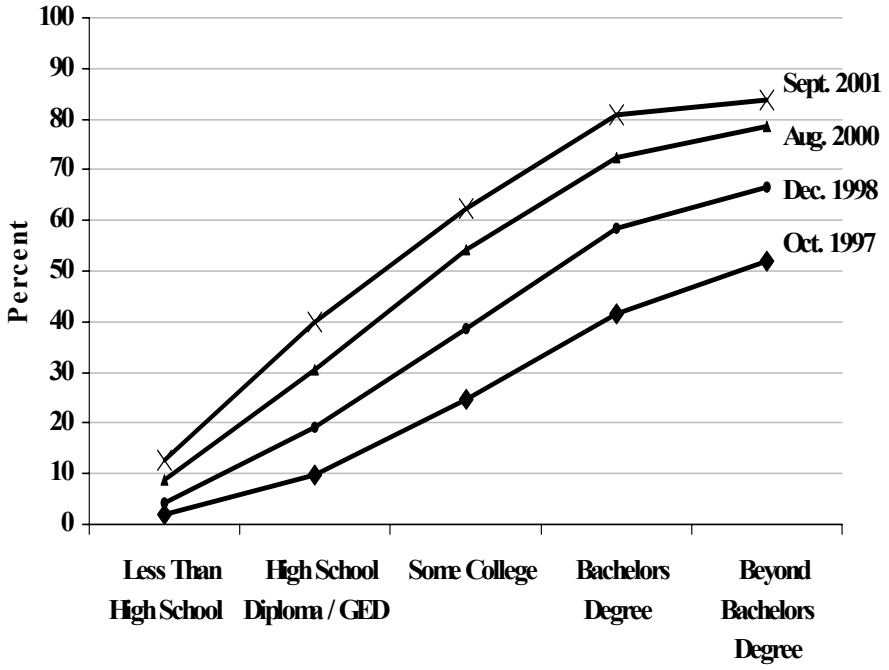


Figure 41: Internet use by education [NTIA02]

[Novak98] also reports that increasing levels of education correspond to an increased likelihood of computer access, regardless of race. However, as depicted in Figure 42, the effects of income and education on Internet use are independent.

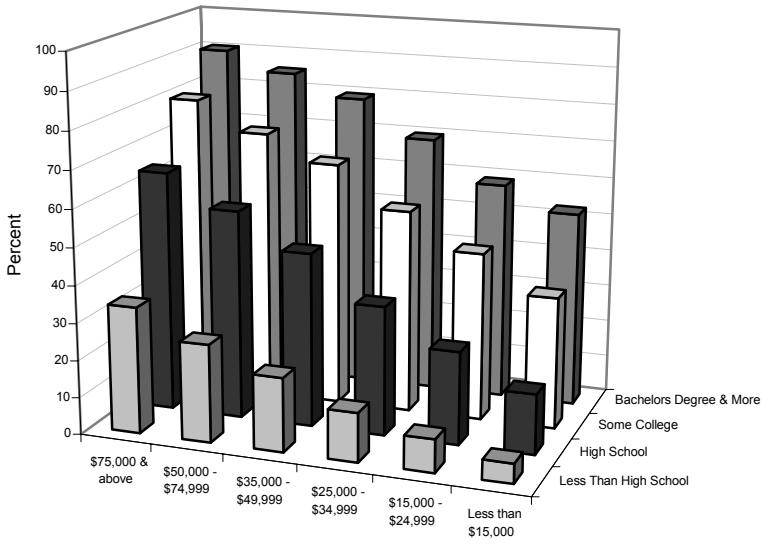


Figure 42: Income and education have independent effects on Internet use [NTIA02]

7.2.5 Divide Based on Region

Rural areas are still lagging behind urban[32] areas (such as the Pittsburgh metropolitan area, where the HomeNet participants live). In 2002, 57.4% of the urban households had Internet access, whereas only 52.9% of the rural households had Internet access. [Novak98a] reports that this lagging behind in Internet access is regardless of income level: "Indeed, at the lowest income levels, those in urban areas are more than twice as likely to have Internet access than those earning the same income in rural areas."

[32] Central city excluded.

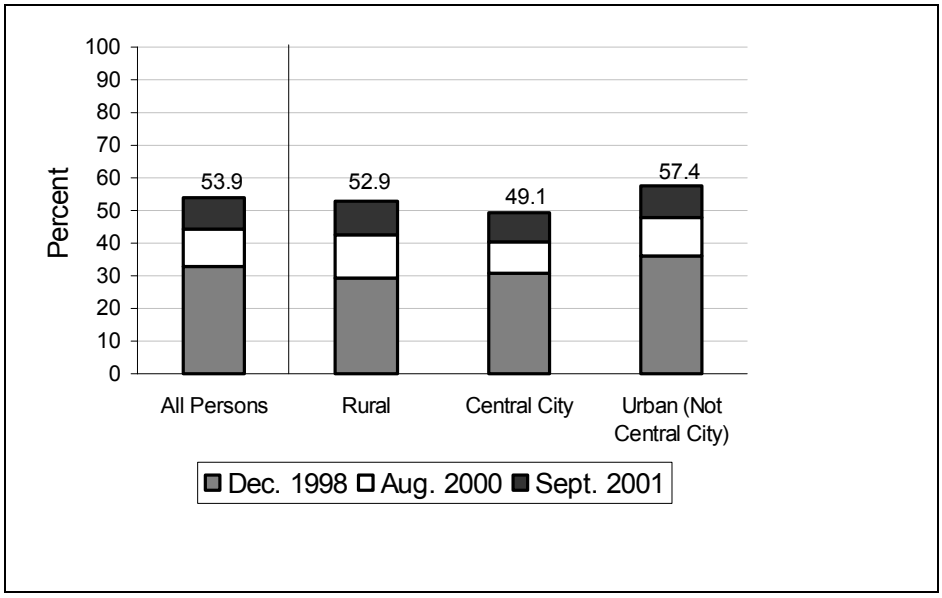


Figure 43: Internet use by geographic location of household [NTIA02]

7.2.6 Divide Based on Age

Computer and Internet use are strongly associated with age. In the last few years, the entire age distribution has shifted upward constantly. Figure 44 and Figure 45 reveal that teenagers are the most likely computer and Internet users. Also, people in their prime workforce years are likely computer and Internet users.

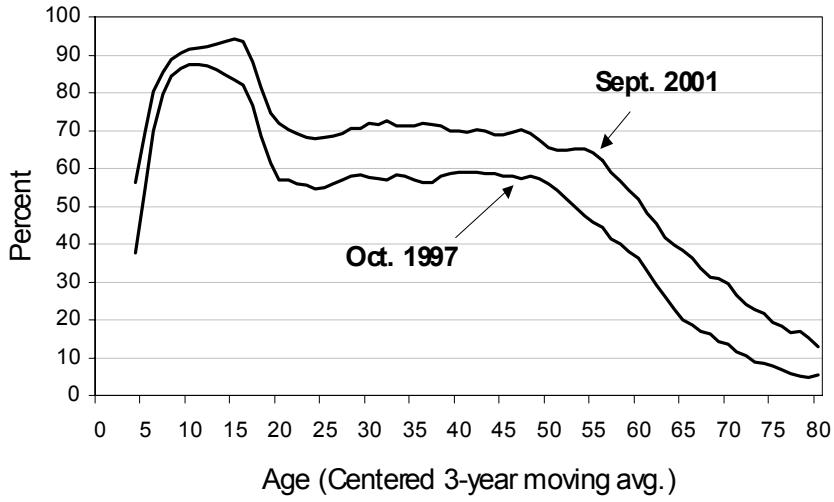


Figure 44: Computer use age distribution (3 year moving average) [NTIA02]

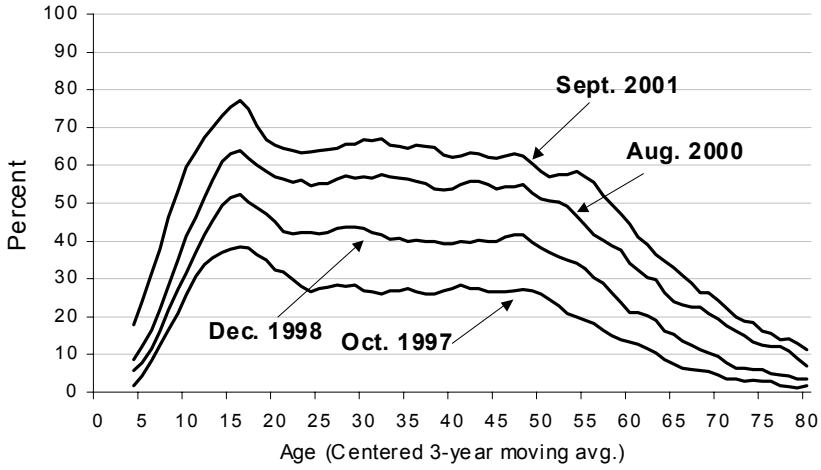


Figure 45: Internet use age distribution (3 year moving average) [NTIA02]

7.2.7 Reasons for Discontinuing Internet Access

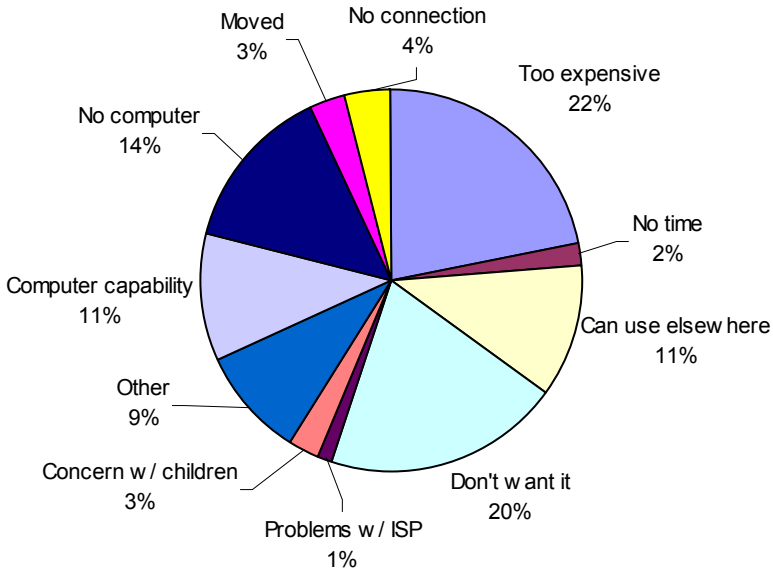


Figure 46: Reasons for United States households discontinuing Internet access percent distribution [NTIA02]

[NTIA02] reports that many of those not online have no intention of going online. Other discontinued Internet use. Among the biggest reasons was ‘don’t want it’ (20%) (See Figure 46). In this regard, a digital divide may not exist only due to economic and other

barriers to usage but also due to the fact that users just don't want Internet access. On the other hand, general cost involved accounted for 22%, which confirms the existence of economic barriers to Internet use.

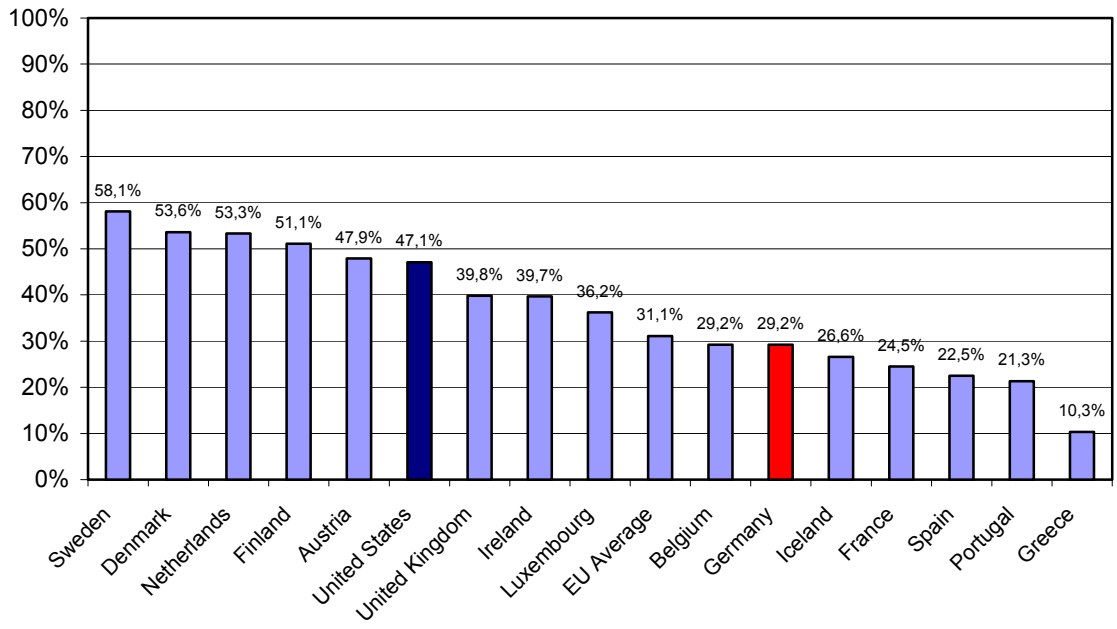
7.3 The Digital Divide in Europe

The rapid diffusion of the Internet is not a unique United States phenomenon. According to OECD data, it is truly a global phenomenon (see Figure 47). Therefore, it is reasonable to extend research on the digital divide in the United States to other countries. This section summarizes the key findings of existing research on the digital divide in Germany.

Although the danger of a social and economical divide in the society is apparent, there is a lack of research on the digital divide in the European and German context. In this regard, the factual foundation for key policy initiatives to promote access to the Internet for all groups of people is still in an introductory stage. Key policy initiatives in Germany include the "Innovation und Arbeitsplätze" initiative and "Initiative D21" [Perillieux00].

This is particularly true with respect to research on Internet usage in residential settings. The importance of measuring the digital divide in residential settings is highlighted by the fact that 73% of the German citizens can use the Internet at Home [Webgauge].

According to a recent study by BoozAllen&Hamilton [Perillieux00], the current situation in Germany can be considered a good start. The telecommunication infrastructure is excellent and PC penetration is high. Unfortunately, Germany is only ranked average with respect to information technology penetration (if compared to other developed countries, see Figure 47). Sweden has the highest Internet usage in the world. However, the United States stays the nation with the highest PC penetration, the highest number of users and the longest duration of online sessions.



Note: EU country estimates are for February 2001 and US estimates are for August 2000.
Source: European Union, http://europa.eu.int/information_society/eeurope/index_en.htm and U.S. Department of Comm
Economic and Statistics Administration.

Figure 47: Individuals using the Internet, selected countries [NTIA02]

According to [Perillieux00], age, level of education, and geography are again the most influencial factors of different Internet usage (see Figure 63, Figure 64, and Figure 65 in appendix B). Tab. 27 reports such a digital divide with respect to age in Germany. Further, a comparison with other European countries reveals that this age gap is rather wide.

Tab. 27: Internet access and age – The digital divide in Europe [Webgauge]

Age	14-29 years	30-49 years	50-69 years
UK	80%	58%	36%
Netherlands	84%	78%	33%
Germany	**67%**	**51%**	**24%**
Belgium	72%	44%	19%
Spain	70%	38%	9%
France	60%	28%	16%

With respect to the dynamics of Internet usage, surveys by Webgauge [Webgauge] and studies by Scholz [Scholz01] report that even though the number of users going online in Germany is still growing, the growth in the number of users is slowly saturating. Also, articles in the German weekly computer magazine 'Computerwoche' [CW01] point out that one third of German citizens are using the Net, but that another third will stay offline [CW02]. This is also confirmed by [Perillieux00] who reports that 21 million Germans between the ages of 14 and 69 will not have the opportunity to use the Internet by 2003. Similar to the situation in the United States, a divide with respect to education and geography is about to widen. Due to better education and easier access, people below the age of 40 will have a significantly higher probability of using the Net.

In summary, it is clearly the case that the typical Internet user in Germany is not representative for the overall German population. In 2000, only 22% of the female population used the Net, whereas 35% of the male population used the Net frequently. Moreover, almost 20% of the Internet users are between age 14 and 19 [Perillieux00]. Moreover, the structure of Web users stays constant over time. Therefore, an existing digital divide is not going to close in the near future.

The effects of this situation on people not using the Internet are similar to the effects the digital divide has in the United States: "these people may face difficulties in finding jobs, shopping online, taking advantage of public or private services, increasing their level of education. Ultimately, these things may result in considerable disadvantages for Germany as an industrial nation" [Perillieux00].

7.4 The Digital Divide from a Global Perspective

The term "digital divide" is often used within a United States-centric context only. However, more than half of the world's population has never made a telephone call. Also, the majority of people in the third world often do not even have access to clean water. Less than 7% of the world's population is online. On the other hand, the United States has 36% of the world's Internet users [NTIA02].

Nielsen/Netratings reports other facts that put the digital divide issue into a global context:

- "The United States has more computers than the rest of the world combined. When assessed by region, Internet use is dominated by North Americans:
- 41% of the global online population is in the United States & Canada

- 27% of the online population lives in Europe, the Middle East and Africa (25% of European Homes are online)
- 20% of the online population log on from Asia Pacific (33% of all Asian Homes are online)
- Only 4% of the world's online population are in South America"

(Source: First Quarter 2001 Global Internet Trends, Nielsen/Netratings)

There exist vast differences in the availability of home Internet access even among highly developed nations. For example, Sweden ranks as the nation with the highest percentage of home Internet connections at 61%, whereas Spain ranks as the nation with the lowest percentage of Internet connections in Europe: only 20% of its homes are connected (see also Figure 47) [NTIA02].

7.5 Implications

7.5.1 General Implications

Information tools, such as personal computers and the Internet, are increasingly critical to economic success and personal advancement [NTIA99b]. Therefore, policymakers must ensure that all people have the information tools and skills that are critical to their participation.

More people than ever have access to telephones, computers, and the Internet. Policymakers in the developed countries have achieved high levels of telephone connectivity through the implementation of two key initiatives: Pro-competition policies and universal service policies. The first have resulted in lower prices for consumers of telephone services. The second have helped assure that most people can enjoy affordable access today.

However, there still exists a significant "digital divide" with respect to Internet usage. This divide has widened as the information "haves" outpace the "have nots" in gaining access to electronic resources [NTIA99b]. The consequences of that are expected to be severe. The Internet may provide equal opportunity and democratic communication, but only for those with access. Such unequal access to the Net will result in a lack of technological skills and will have a severe impact on the competitiveness of a country's firms. Therefore, until every home can afford access to information resources, there will be a need for public policies and private initiatives to expand affordable access to those resources.

People who are unable to participate in electronic commerce cannot benefit from the convenience and tax savings of online shopping in the United States and elsewhere. Businesses with no Internet presences are at a competitive disadvantage to businesses with such presence because they cannot offer tax-free merchandise. Other examples that clarify the disadvantage of people who do not have access to the Internet include:

- Even after the burst of the dot.com bubble, employers increasingly look especially for Internet skills. Chances of the Job market for people without access are decreasing. In this regard, the attractiveness of a region for investors is highly dependant on the level of education of its people, especially with respect to Internet skills. Also, applying online is not possible for those who do not have access. Unfortunately, unemployed citizens are not the typical users of the Net.

- Companies are increasingly offering low-price online-only services. Examples include Internet banking and rebates for rental cars ordered online. People without access to the Net will not be able to take advantage of these services. They will have to use the often more expensive offline service.

- Access to the Internet is an advantage for students who do research for a particular homework.

- Public services will increasingly offer services online. People without access to the Internet will have to use the time-consuming offline services.

The reasons for discontinuing Internet use are a first step towards addressing the problem of the digital divide. For example, cost ranked highest among reasons for discontinued Internet use (see Figure 46). To bridge the gap, there are local and national efforts to distribute hardware and software to those in need. Assistance for low-income households and support for high-cost regions are prime examples of such programs. According to [Novak98], policymakers should ensure participation for all people in the Information revolution by improving the educational opportunities for minorities. Also, policymakers should aim to increase telephone penetration and to expand computer and Internet connectivity even further, particularly among the underserved. [NTIA99b] lists lower prices, leasing arrangements, and free computer deals as actions that would help to bridge the digital gap. In all these discussion of the digital divide, the usual assumption has been that access to the Web automatically translates into usage and thereby helps close the digital divide. If access is ensured (e.g., by offering training and reducing the cost of access), usage will automatically follow. In the HomeNet sample, since all users received hardware, Internet connection, and even training, the penetration level is already 100%. The question that this dissertation and other studies answer is whether or not we

are still seeing a digital divide with respect to the level of use of this technology. One of the key findings of this dissertation is that even after removing economic, educational and technological barriers to usage, the fact remains that not all groups of people use the Internet. As discussed previously, e.g., in Section 3.3.3, there are race and gender differences with respect to Internet usage in the HomeNet sample. These findings imply that increased utilization of the Web will require more than access. Therefore, even though the most important and first steps to narrowing the divide are providing access to the technology and the availability of infrastructure to facilitate its use, access alone is not enough to address the problem. Numerous other factors come into play.

Other reasons for not using the Internet include 'concerns with children', which accounted for 3% of the reasons. Therefore, policymakers have to ensure that the technology and educational resources are given to parents to protect children from inappropriate material.

Another issue that quickly gains importance (even though it is not yet shown in Figure 46) is how privacy concerns should be addressed. Many people are concerned about invasion of their privacy. According to [NTIA99b] almost 2/3 of Americans are either "very concerned" or "somewhat concerned" about confidentiality on the Internet. There are private Sector initiatives that deal with this issue, e.g., BBBOnline [BBBOnline] and TRUSTe [TRUSTe], which require merchants to adhere to fair trade practices.

One of the primary policy goals should be to have the Internet become a mass-market phenomenon. Mobile access and broadband access will be among the key enablers [Perillieux00]. However, with the burst of the dot-com bubble, it clearly has become more difficult from a financial point of view to afford broadband access. Also, many firms with business plans that involved mobile technology went bankrupt during the crisis.

In summary, universal service policies will remain of critical importance. However, there are also language and other cultural barriers to be addressed. The content of Web sites still does not address all groups of people. Also it is important to build awareness among people who do not realize that this technology is relevant to their lives.

7.5.2 Specific Implications for Germany

As in the United States, unequal access to the Internet will result in a lack of technological skills for subgroups of people, which will clearly have an impact on the competitiveness of German firms. In order to avoid a digital divide in Germany, we can learn from studies that have been conducted in the United States. However, some important differences prevail.

For example, tax reductions for people shopping online are no argument against the digital divide, because they do not exist in Germany.

The United States was also the first country that started dealing with the digital divide issue in 1995 and is measuring its extent ever since. Germany started to deal with the digital divide comparatively late. Government reports on the issue, which are comparable to the series of reports published by the United States government (NTIAa, NTIA95, NTIA99b, NTIA02]) are still very rare.

In this regard, the following key requirements must be met [Perillieux00]:

- Building of a data basis for research on the digital divide, similar to the 'Falling through the Net' report in the United States
- Setting of explicit goals with regard to the further development of the information infrastructure, which should be easily quantitatively measurable, such as Internet penetration.
- Transformation of non-users to users by targeting drivers and impediments of non-usage, such as usage cost, Internet skills, availability of attractive content and applications in the Net.

In general, it is important to continue to build a telecommunication infrastructure, offer education, to adapt legal issues to the new e-environment, and to offer financial incentives to firms. In this regard, private and public initiatives in the United Kingdom and the United States may be considered best practices, although adaptation to German peculiarities will be necessary. For example, the federal nature of Germany will make a cooperation of federal government, state government, public institutions, and private institutions necessary.

In this regard, the private and the public sector (at both, the state and the federal level) should work together to avoid and to overcome a possible digital divide. Incentive mechanisms for the private sector will be of utmost importance. One example of such cooperation in Germany is the initiative "Schulen ans Netz" [SAN], which aims to increase the number of Internet access points in German schools. Access points in schools and public institutions are considered a key enabler for bridging the digital divide. In 1999, United States schools received 30 times as much money for Internet access than German schools [Perillieux00].

The initiative of the German government 'Innovation und Arbeitskräfte' has also the goal to increase Internet usage and the penetration of modern information and communication

technology. This initiative aims to ease access for the predicted 21 million German non users in 2003 and addresses a variety of barriers to usage:

- Cost of access (see Figure 68 in appendix B)
- Lack of transparency with regard to the variety of initiatives and no central coordination of these initiatives due to the federal nature of the German system
- Internet skills and ease of usage
- Content

The German government is an important driver of Internet technology. However, its power is limited. For example, the digital signature is on the market since 1997, however, it is seldom used. The leading role of the state as an 'early adopter' could encourage other institutions to follow. For example, there is a need for cost reduction and streamlining of processes of public services. Therefore, public services are increasingly offered online. However, because a significant part of the German population does not use the Internet, state institutions will be forced to provide its service for both, Internet users and people who do not use the Net. In the beginning, this will lead to even higher cost instead of cost reduction [Perillieux00].

8 Concluding Remarks and Future Work

Using IT in general and the Internet in particular is expected to increase the efficiency of economies. Customer relationships are transforming. The transparency of the market and consumer power is increasing [Staudt01]. In such a world, competition is only one click away. Many companies already took the first step towards the online world [Bensberg01]. For these companies, it is important to know who is online. Research on individual Web usage is a necessary precondition for successful electronic commerce, but also for successful key policy initiatives that ensure equal access to the Internet for all groups of people. In this regard, the presented results have important implications both for business-to-consumer electronic commerce and for public policy as it pertains to the digital divide. Also, the knowledge of regularities, such as the saturation levels of Web usage and the power law distribution of Web site popularity, helps reveal the historical evolution and future behavior.

The interrelated studies presented in this dissertation advance the research on Web usage in several aspects. They measure, analyze, and interpret individual Web utilization and Web loyalty using a variety of metrics. By taking repeated measures of Web usage over time, these studies identify significant trends in Web usage. Several findings are surprising. The scientific work presented in this dissertation advances the research on Web usage by:

- exploring whether the increase in Web site visiting opportunities spurred an increase in the Web utilization rates of individual users (see chapter 3: ‘Saturation of Lay Web Usage’ on pp. 40 ff.),
- applying session-based measures to gain insights in individual Web usage in Web sessions and identify how Web users change the way they use the Web as their individual level of expertise increases (see chapter 4: ‘Analyzing Web Sessions’ on pp. 62 ff.),
- addressing the question whether different user groups also differ in loyalty to Web sites and whether users converge over time to a set of ‘favorite’ Web sites (see chapter 5: ‘Web User Loyalty and Web Site Stickiness’ on pp. 80 ff.),
- specifically addressing the issue of Web portal utilization to answer the question whether Web portal users are different from average Web users (see chapter 6: ‘Portal Utilization’ on pp. 99 ff.).

The major results of this dissertation with implications for *electronic commerce* are as follows:

- Web usage is not distributed equally across subgroups of users. Web users can be clustered into four groups with distinct trajectories of Web usage. All groups reach saturation in their extent of Web usage after following a downward path. These saturation levels turned out to be comparatively low for most users, only a few users utilized the Web heavily. We observe saturation of Web usage independent of the specific measure of distinct Web sites visited. Surprisingly, there are no material differences between trajectory groups in terms of their utilization intensity as measured in page views per site. Individual characteristics, particularly gender and ethnic background, determine the saturation levels of Web usage. Minorities and females utilize the Web to a lesser extent.
- Web users spent only limited time in the Web. Regardless of the measurement of Web usage in Web sessions used, only a small group of users uses the Web heavily. We identified characteristics of individuals that influence Web usage in Web sessions, which include ethnic background, gender, household income, phone usage, e-mail usage, and computer skill level. In particular, belonging to a minority group and being female determines low Web usage in Web sessions, which confirms the findings from the previous chapter. Surprisingly, there does not seem to be a significant shift from undirected browsing to directed access of Web sites over time. This seems to confirm that users keep exploring that Web even after months of Internet experience.
- Given limited capacity for Web utilization, sites that can achieve high rates of repeat visits (and purchases) are likely at a clear advantage. Therefore, issues such as loyalty in the Web and stickiness of Web sites are increasingly crucial for marketing in the Web. We reveal that users show little loyalty to Web sites. There is considerable churn in Web sites visited across subgroups of users. The degree of churn is a constant over time across all groups of users. Therefore, measures of Web site success such as popularity and stickiness are needed. We reveal that Web sites differ substantially in popularity and stickiness. In particular, only a few Web sites were popular with most users. Among the popular Web sites, yahoo.com dominates the other sites with respect to popularity and stickiness. It is both able to acquire and retain customers.
- With respect to Web portal utilization at a specific portal, yahoo.com, only 23% of the users visit more than one of the services offered, only 13% visit more than two services. The ethnic background of individuals has no significant impact on portal utilization in the World Wide Web! This is an important and surprising result given the

fact that the ethnic background does have a significant impact on *overall* Web utilization. It seems that portal sites speak more to minorities and women than other sites in the Web.

The results predict some fundamental consequences. Some of these predictions have already become reality. For example, all the presented results predict that the Web is highly competitive and the degree of competition is increasing even further. The Web can be thought of as a marketplace with sites competing to attract users to visit. The saturation levels for Web site visits in every trajectory group identified in chapter 3 can be interpreted to estimate the size of this market and help estimate the number of potential Web site visiting opportunities that Web sites will compete for each month. As the number of new Web users begins to decrease, the number of Web site visiting opportunities will reach a steady state. Moreover, competition among Web sites for these visiting opportunities seems to be a zero sum game and grows even further in intensity. The presented results speak also for a more critical contemplation of the effectiveness of some business models in the Web, particularly those that rely on page impressions.

High churn or low loyalty of customers in the field of e-commerce is becoming a severe problem in e-commerce [Eifert01]. The reason for that is decreasing information cost for consumers [Bakos97], increasing transparency of the market und easier churn of customers. On the other hand, decreasing information cost lead to information overload. Our finding of extraordinary high and constant churn in Web sites visited across subgroups confirms [Bakos97]. Given a degree of churn greater than zero, limited capacity of users turns competition into a zero sum game for competitors in the World Wide Web.

The relation between stickiness and popularity gives us insights into the success or failure of some Web sites with low stickiness when the growth of the Internet in number of new users slows down or even stops. In this regard, this dissertation provides some guidelines on the success or failure of online businesses. In the HomeNet data, only a few sites were broadly popular. Moreover, only a few succeed in retaining visitors. The disproportionate distribution of user volume among sites is characteristic of winner-take-all markets, in which the top few contenders capture significant market share.

One of the measures taken by Web sites to increase loyalty is to offer additional services. Web portals have implemented a variety of services to attract and retain visitors with varying degrees of success. Indeed, they invested substantial amounts of money in these additional services. It is interesting to examine how this investment paid off. The low utilization rates we observed may be due to marketing problems of yahoo.com.

In general, analyzing the characteristics of Web users online is very appealing from a marketing perspective. Specifically, linking Web usage data with, e.g., demographic data, as conducted in this dissertation, allows for creating user profiles that is the basis for targeted marketing. The key findings of this study include determinants of Web usage measured by various metrics, such as monthly visits, session-based Web usage, and Web portal utilization. Such knowledge on user characteristics can be used by companies that want to do business online in general and by marketing departments for targeted marketing in particular. As expected, the typical Web user in 1995-1997 does not represent the average citizen, even though the data sample consisted of average citizens.

Apart from the fact that the identification of factors that characterize the different user groups provides valuable insights from a business perspective, the results reported in this study also have importance from a *public policy* perspective. The question that this dissertation addresses is whether or not we are seeing a digital divide in the HomeNet data with respect to the level of use of Internet technology. As expected, many of the findings confirm other studies on the digital divide. However, many other findings presented in this dissertation lead to new conclusions.

The major results of this dissertation with implications for *public policy* are as follows:

- With regard to monthly Web utilization measured by the number of distinct Web sites visited and Web utilization intensity measured by the number of page views, there are indeed race and gender differences with respect to Web utilization in the trajectory groups. Whites use the Web significantly more than people who belong to a minority group. Similar differences in the utilization of the Web by gender can be observed. Heavy and very heavy users of the Web tend to be male, whereas light users tend to be female. Also, there is a digital divide with respect to age. Heavy and very heavy users of the Web tend to be younger than light and moderate Web users.
- Analyzing individual Web usage in Web sessions reveals that, in contrast to other studies on the digital divide, household income does have a negative impact on Web utilization. This is in contrast to previous findings of HomeNet, which say that income did not predict Internet use. Gender, race, and generation were all strong predictors of Internet use in the sample. Surprisingly similar, age has a positive impact on Web usage in sessions. Prior work on the HomeNet data, such as described in chapter 3, did find a negative impact of age on Web use. Use of a more subtle session-based approach for measuring usage has advanced this work and lead to new conclusions.
- With regard to Web portal utilization, this dissertation aims at examining whether a possible digital divide exists with respect to portal utilization to the same extent as it is

proven to exist with respect to overall and session-based Web usage. Even though there still existes a gender gap, the identified gender gap is much smaller and the race gap just does not exist. Moreover, heavy portal users do not share the same demographic characteristics with people with high overall Web utilization. There is less of a digital divide with respect to portal usage than with respect to overall Web usage. It seems that portal sites speak more to minorities and women than other sites in the Web. Apart from the slight impact of age and gender on usage, the demographic characteristics are surprisingly similar across the groups with different portal utilization. Moreover, in contrast to the results of studies on the digital divide, which show that heavy and very heavy users tend to be younger, age even has a positive impact on portal usage in this analysis.

In general, this dissertation advances the research on the digital divide by addressing the question of whether access to the Web automatically translates into Web usage. As noted in chapter 7, it is commonly acknowledged that access translates into usage. The HomeNet data set allows for direct testing of this hypothesis. As discussed in chapter 2, all the users in the HomeNet sample received hardware, Internet connection, and basic training for no charge. However, even though the people had access for free, there were still subgroups of people not using the Web to the same extent. These findings imply that increased utilization of the Web will require more than access. The problem is not simply one having access, but rather one having the capacity plus the interest in using this technology. Following this sequence, at the time the HomeNet data was collected, the information on the Web was probably more appealing to whites than minorities. Therefore, more online information that is appealing to the minority community is needed. Informal reports indicate that customized training by gender or race may be needed in addition to access to enable different segments of society to benefit fully from the Internet. Even though recent studies show that the gender gap is closing in terms of time spent online, men & women use the Internet different in terms of services used. In this regard, additional work is required to develop policies that will be more successful in promoting utilization of the Web.

According to the literature review presented in chapter 7, the digital divide identified in the HomeNet data does still exist and is in many cases even widening. Therefore, research on the digital divide will continue to be a very important issue in the new millennium. These findings do not signal the end of new findings about Web usage, but do establish a firm foundation to build upon in further research.

Specific **future work** issues in terms of *data collection* and in the direction of *methodological extenions* have already been discussed in the previous chapters of this dissertation. More general issues are described below:

With regard to *data collection* issues, we have already emphasized the value of more recent data. We particularly emphasize the value of studies like the present one for German marketing specialists and policy makers. Therefore, especially in Germany, further research on Web usage in general and the digital divide in particular is desirable. However, there is no comprehensive and valid data basis for such research in Germany. With respect to research on the digital divide, not much data is available on minority groups (eg., the Turkish or Italian people). The series of reports on the digital divide in the United States can be considered best practice that should be conducted – with some modifications – in Germany too. However, the comparatively strict German privacy regulations impede building up such a data basis. For example, the German 'IuKDG'[33] says that using and analyzing individual Web usage data is only possible if the user explicitly allows that. Storing individual Web usage data without this permission is illegal (see also [Höller99]). Moreover, linking Web usage data with, e.g., demographic data, which allows for creating user profiles as a basis for targeted marketing, is difficult. According to the German 'Persönlichkeitsgesetz', making this link between usage and demographic data is also illegal (see also [Engel97]). This holds even if a given individual gives his explicit permission to link his demographic data with this usage data [Wittig00]. These data protection and privacy regulations in Germany must be considered a severe restriction of research and marketing activities. Some software companies that do business in the field of logfile analysis even consider it a non-tariff barrier. Therefore, further studies in the German context within the boundaries of privacy regulations are strongly encouraged. Further, such future work does not have to be limited to Web usage from fixed access points. Because of the growing importance of wireless Internet access, it would be desirable to analyze individual usage of data services on the cellular networks in Europe. This would shed light on mobile access to the Internet - which in Europe is more important than in the States at the present time.

With respect to the *methodological direction*, it would be interesting to investigate what other methodological approaches (e.g., latent variable analysis) are appropriate for the kind of analysis conducted in this dissertation. This includes methodological approaches

[33] Informations- und Kommunikationsdienstegesetz

that help identifying overall Web browsing trends in general and Web loyalty, portal utilization, and session identification in particular.

The Internet today is only the basis for things to come. It is possible that the limited capacity for Web site visits – as reported in this dissertation - is due to the current technical shortcomings on the Internet (e.g., ease of use of sites, difficulty in using search engines, ineffectiveness of banner advertisements). Breakthroughs in technology can potentially increase the capacity for Web utilization and in turn the size of the market. The future of the Net would probably be wireless [Staudt01]. Sales of common personal computers are already saturating. Pervasive computing, processors, memory chips, network hardware will be small and affordable mass products. New devices such as PDAs will make access more ubiquitous [Staudt01]. As the Internet becomes a more pervasive technology, further studies are needed to cope with the forthcoming changes.

References

Abrams97a — Abrams, A., Diversity and the Internet, Journal of Commerce, 1997, *June 26th*

Accrue — http://www.accrue.com/ [5-Nov-02]

Adamic01 — Adamic, L. A., Huberman, B. A., The Web's Hidden Order, Communications of the ACM, 2001, *9*, ACM Press, 2001, p. 55-60, http://doi.acm.org/10.1145/383694.383707 [5-Nov-02]

ADO — http://www.microsoft.com/data/ado/default.htm [5-Nov-02]

Agrawal01 — Agrawal, V., Arjona, L. D., Lemmens, R., E-performance: the path to rational exuberance, McKinsey Quarterly, 2001, *1*, p. 31-43

Anderson95 — Anderson, R., Bikson, T., Law, S., Bridger, M., Universal Access to E-Mail: Feasibility and Societal Implications, Santa Monica, CA, Rand Corporation, 1995

Bakos97 — Bakos, Y., Reducing Buyer Search Costs: Implications for Electronic Marketplaces, Management Science, 1997, *12*, p. 1676-1692

Bakos98 — Bakos, Y., The Emerging Role of Electronic Marketplaces on the Internet, Communications of the ACM Business Review, 1998, *8*, p. 35-42

BBBOnline — http://www.bbbonline.org [5-Nov-02]

Bensberg01 — Bensberg, F., Warenkorbanalyse im Online-Handel, Buhl, H. U., Huther, A., Reitwiesner, B.: Information Age Economy. 5. Internationale Tagung Wirtschaftsinformatik, Heidelberg, Physica-Verlag, 2001, p. 103-116

Bensberg99 — Bensberg, F., Weiß, T., Web Log Mining als Marktforschungsinstrument für das World Wide Web, Wirtschaftsinformatik, 1999, *5*, p.426-432

Berendt01 — Berendt, B., Mobasher, B., Spiliopoulou, Measuring the Accuracy of Sessionizers for Web Usage Analysis, Proc. Workshop on Web Mining at the First SIAM International Conference on Data Mining, 2001, p. 7-14

Berendt02a Berendt, B., Mobasher, B.Nakagawa, M., Spiliopoulou, M., The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis, Hrsg.: Masand, B., Spiliopoulou, M., Srivastava, J., Zaiane, O., Working Notes of the Fourth WebKDD Web Mining for Usage Patterns & User Profiles Workshop at KDD 2002, Alberta, CA, 2002, p. 115-129

Berendt02b Berendt, B., Hotho, A., Stumme, G., Towards Semantic Web Mining, Hrsg.: Horrocks, I., Hendler, J., The Semantic Web - Proceedings of the 1st International Semantic Web Conference, Sardinia, Italy, Springer, 2002, p. 264-278

Boneva01 Boneva, B., Kraut, R., & Frohlich, D., Using e-mail for personal relationships: The difference gender makes, American Behavioral Scientist Special Issue: The Internet in everyday life, 2001, *3*, p. 530-549

Cameron86 Cameron, A. C., Trivedi, P. K., Econometric models based on count data; comparisons and applications of some estimators and tests, Journal of Applied Econometrics, 1986, *1*, p. 19-53

Catledge95 Catledge, L., Pitkow, J., Characterizing browsing strategies in the world wide web, Computer Systems and ISDN Systems, 1995, *6*, 1995, p. 1065-1073

Chau99 Chau, M. Y., Web Mining Technology and Academic Librarianship: Human-Machine Connections for the twenty-First Century, First Monday, 1999, *6*

Choo00 Choo, C. W., Detlor, B., Turnbull, D., Information Seeking on the Web: An Integrated Model of Browsing and Searching, First Monday, 2000, *2*

Chow97 Chow, S., Reed, H., Toward an Understanding of Loyalty: The Moderating Role of Trust, Journal of Managerial Issues, 1997, *3*, p. 275-298

Christ01a Christ, M., Krishnan, R., Nagin, D., Kraut, R., Günther, O., Trajectories of individual WWW usage: implications for electronic commerce, Proc. 34th Hawaii International Conference on System Science (HICSS-34), IEEE Press, 2001

Christ02a Christ, M., Baron, S., Krishnan, R., Nagin, S., Günther, O., Advancing Measurements of Web Usage: A Session-Based Approach, Working Paper Submitted to the 11th European Conference on Information Systems, 2002

Christ02a Christ, M., Krishnan, R., Nagin, D., Günther, O., Measuring Web Portal Utilization, Proc. 35th Hawaiian Conference on Information Systems, IEEE Press, 2002

Christ02b Christ, M., Krishnan, R., Nagin, D., Günther, O., An Empirical Analysis of Web Site Stickiness, Proc. 10th European Conference on Information Systems, Gdansk, 2002

Christ02c Christ, M., Krishnan, R., Nagin, S., Günther, O., Saturation of Lay Internet Usage - Implications for Electronic Commerce and Public Policy, Working Paper - Submitted to Management Science, 2002

Clark96 Clark, K., Kalin, S., Techno-stressed Out? How to Cope in the Digital Age, Library Journal, 1996, *13*, p. 30-35

CLF http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format [5-Nov-02]

CLS The Common Logfile Format, http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format [5-Nov-02]

Cooley00 Cooley, R., Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data - PhD Thesis, University of Minnesota, 2000

Cooley99 Cooley, R., Mobasher, B., Srivastava, J., Data Preparation for Mining World Wide Web Browsing, Knowledge and Information Systems, 1999, *1,* p. 5-32

Cummings02 Cummings, J. N., & Kraut, R., Domesticating computers and the Internet, The Information Society, 2002, *3*, p. 221-232

Cutler00 Cutler, M., Sterne, J., E-Metrics: Business Metrics For The New Economy, Cambridge, MA, 2001, http://www.netgen.com/emetrics [5-Nov-02]

CW01 Jeder dritte Deutsche ist "drin", Computerwoche, 2001, *Feb 9th*

CW02 Jeder dritte Deutsche bleibt offline, Computerwoche, 2002, *June 16th*

Cyb98a CyberAtlas. Demographics: Who's on the Net in the US?, http://www.cyberatlas.com/demographics.html [5-Nov-02]

DDN The Digital Divide Network, http://www.digitaldividenetwork.org [5-Nov-02]

Double96 DoubleClick Frequency research findings, July 1996, http://www.doubleclick.net/us/resource-center/findings/banner-burnout.asp [22-Apr-00]

Doyle99 Doyle, R,, Privacy in the Workplace, Scientific American, 1999, *1*, p. 19

DUnger98 D'Unger, A., Land, K., McCall, P., & Nagin, D., How many latent classes of deliquent/criminal careers? Results from mixed Poission regression analyses of the London, Philadelphia, and Racine cohorts studies, American Journal of Sociology, 1998, *103*, p. 1593-1630

Economist00 E-Commerce: Shopping Around The World, The Economist, 2000, *Feb. 26th*, p. 5-54

Eifert01 Eifert, D., Pippow, I., Erfolgswirkungen von One-to-One Marketing - Eine empirische Analyse, Buhl, H. U., Huther, A., Reitwiesner, B.: Information Age Economy. 5. Internationale Tagung Wirtschaftsinformatik, Heidelberg, Physica-Verlag, 2001, p. 265-278

Engel97 Engel-Flechsig, S., Teledienstedatenschutz - Die Konzeption des Datenschutzes im Entwurf des Informations- und Kommunikationsdienstegesetzes des Bundes, Datenschutz und Datensicherheit, 1997, *1*, p. 8-16

Exody http://www.exody.net/ [5-Nov-02]

Fenichel81 Fenichel, C. H., Online searching: measures that discriminate among users with different types of experiences, Journal of the American Society for Information Science, 1981, *1*, p. 23-32

Ferguson01 Ferguson, G. T., Pike, T. H., Mobile Commerce: Cutting Loose - Making a Shift from m to u, Buhl, H. U., Huther, A., Reitwiesner, B.: Information Age Economy. 5. Internationale Tagung Wirtschaftsinformatik, Heidelberg, Physica-Verlag, 2001, p. 7-13

Fidel98 Fidel, R., Davies, R. K., Douglas, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K., Toney, C. D., A visit to the information mall: Web searching behavior of high school students, Journal of the American Society for Information Science, 1998, *1*, p. 24-37

Frank95 Frank, R., Cook, P., The Winner-Take-All Society, New York, The Free Press, 1995

Fukuyama95 Fukuyama, F., Trust: The Social Virtues and the Creation of Prosperity, New York, The Free Press, 1995

Ganzel98 Ganzel, R., Feeling Squeezed by Technology?, Training, 1998, *4*, p. 62-70

Garfinkel99 Garfinkel, S., Spafford, G., Web Security & Commerce, Cambridge, O'Reilly & Associates, 1997

Gefen02 Gefen, D., Customer Loyalty in E-Commerce, Journal of the Association for Information Systems, 2002, *3*, p. 27-5

Gomory99 Gomory, S., Hoch, R., Lee, J., Podlaseck, M., Schonberg, E., Analysis and Visualization of Metrics for Online Merchandizing, Proc. WEBKDD Workshop of Web Usage Analysis and User Profiling, 1999

Görsch00 Görsch, D., Christ, M., Recommender Systems for Electronic Grocery Shopping, Proc. Worshop on Perspectives in Business Informatics Research, Rostock, 2000

Gosh85 Ghosh, J. K and Sen, P. K., On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results, Hrsg.: LeCam, L. M., Olshen, R. A., Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer, Volume2, Monterey, Wadsworth, 1985, p. 789-806

Grimes01 Grimes, A., Closing the Gap, The Wall Street Journal, 2001, *Oct. 29*

Hansell01a Hansell, S., Red Face for the Internet's Blue Chip, The New York Times, 2001, *Nov. 3*

Heskett94 Heskett, J. L., Jones, T. O., Loveman, G. W., Sasser, W. E. J., Schlesinger, L. A., Putting the Service-Profit Chain to Work, Harvard Business Review, 1994, *2*, p. 164-174

HNMAP http://homenet.hcii.cs.cmu.edu/progress/hnmap.html [5-Nov-02]

Hoffman96a Hoffman, D. L., William, D., Kalsbeek, D., Novak, T. P., Internet and Web use in the United States: Baselines for Commercial Development, Communications of the ACM, 1996, *12*, p. 36-46

Hoffman98a Hoffman, D. L., Novak, T. P., Bridging the racial divide on the Internet, Science, 1998, *Apr. 17*

Höller99 Höller, J., Die rechtlichen Rahmenbedingungen des Electronic Business, Hrsg.: Höller, J., Pils, M., Zlabiger, R., Internet und Intranet - Auf dem Weg zum Electronic Business, 2, Berlin, Springer, 1999, p. 351-384

HomeNet The HomeNet project, http://homenet.andrew.cmu.edu/progress [5-Nov-02]

HomeNet95 HomeNet: A Field Trial of Residential Internet Services - Research report, http://homenet.hcii.cs.cmu.edu/progress/report1.html [5-Nov-02]

Horrigan02 Horrigan, J. B., Lee Rainie, D., Getting Serious Online (Report Internet & American Life Project), Washington, DC

Hsieh93 Hsieh-Yee, I., Effects of search experience and subject knowledge of the search tactics of novice and experienced searchers, Journal of the American Society for Information Science, 1993, *3*, p. 161-174

HTTP Gettys, J. M., Frystyk, L., Masinter, P. L., Berners-Lee, T., Hypertext Transfer Protocol - http/1.1. RFC 2616, http://www.w3.org/Protocols/rfc2616/rfc2616.html [5-Nov-02]

IAB Internet Advertising Bureau, Measuring Sucess, http://www.iab.net/measuringsuccess/index.html [5-Nov-02]

IDS00 Internet Domain Survey, Jan. 2000, http://www.isc.org/ds/WWW-200001/report.html [5-Nov-02]

InternetReport Computer Industry Almanac Inc., Internet Report, http://www.c-i-a.com/200103iu.htm [5-Nov-02]

Janetzko99 Janetzko, D., Surfer im Visier, c't, 1999, *20*, p. 86-92

JDBC http://java.sun.com/products/jdbc/ [5-Nov-02]

Jones00a Jones, B. L., Rafaeli, S., Time to Split, Virtually: Discourse Architecture and Community Building Create Vibrant Virtual Publics, Proc. 33rd Hawaii Conference on Information Systems, IEEE Press, 2000

Jones00b Jones, B. L., & Rafaeli, S., What Do Virtual ” Tells” Tell ? Placing Cybersociety Research Into a Hierarchy of Social Explanation, Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE Press, 2000

Jones02 Jones, B.L., Nagin, D.S., & Roeder, K., A SAS procedure based on mixture models for estimating developmental trajectories, Sociological Research Methods, 2001, *29*, p. 374-393

Kass95 Kass, R.E., & Raftery, A.E., Bayes factor, Journal of the American Statistical Association, 1995, *190*, p. 773-795

Keribin97 Keribin, C., Consistent Estimation of the Order of Mixture Models - Working Paper, Laboratoire Analyse et Probabilite, Universite d’Evry-Val d’Essonne, 1997

Khan98 Khan, K., Locatis, C,, Searching thourgh cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web, Journal of the American Society for Information Science, 1998, *2*, p. 176-182

Klawe95 Klawe, M, Levenson, N., Women in Computing: Where are We Now?, Communications of the ACM, 1995, *1*, p. 29-44

Kohavi01 Kohavi, R., Mining E-Commerce Data: The Good, the Bad, and the Ugly, Proc. Knowledge Discovery in Databases, San Francisco, CA, ACM, 2001, p. 8-12

Köhntopp00 Köhntopp, M., Köhntopp, K., Datenspuren im Internet, Computer und Recht, 2000, *4*, p. 248-257

Kraut Kraut, R. E., Lundmark, V., Kiesler, S., Scherlis, W, Mukhopadhyay, J., Why People Use the Internet, 2001, http://homenet.hcii.cs.cmu.edu/progress/purpose.html [5-Nov-02]

Kraut96a Kraut, R. E., Scherlis, W, Mukhopadhyay, T., Manning, J., Kiesler, S., The HomeNet field trial of residential Internet services, Communications of the ACM, 1996, *12*, p. 55-63

Kraut96b Kraut, R. E., Scherlis, W, Mukhopadhyay, T., Manning, J., Kiesler, S., HomeNet: A Field Trial of Residential Internet Services, Proceedings of CHI96, 1996

Kraut99 Kraut, R., Mukhopadhyay, T., Szczypula, J., Kiesler, S., Scherlis, B., Information and communication: Alternative uses of the Internet in households, Information Systems Research Special Issue, 1999, *4*, p. 287-303

Land96 Land, K., & Nagin, D., Micro-models of criminal careers: A synthesis of the criminal careers and life course approaches via semiparametric mixed poisson models with empirical applications, Journal of Quantitative Criminology, 1996, *12*, p. 163-191

Mayo33 Mayo, E., The human problems of an industrial civilization, Cambridge, MA, Harvard University Press, 1933

McKenzie01a McKenzie, B., Cockburn, A., An empirical analysis of web Page revisitation, Proc. 34th Hawaii International Conference on System Science (HICSS-34), IEEE Press, 2001

Miller01 Miller, A., Reaching Across the Divide: The Challenges of Using the Internet to Bridge Disparities in Access to Information, First Monday, 2001, *10*

Miller56 Miller, G. A., Miller, G. A., “The magical number seven, plus or minus two: Some limits on our capacity to process information, Psychological Review, 1956, *53*, p. 81-97

Mobasher01 Mobasher, B., Honghua, D., Tao, L., Miki, N., Effective Personalization Based on Association Rule Discovery from Web Usage Data, Web Information and Data Management, Atlanta, GA, 2001, p. 9-15

Montgomery00 Montgomery, A. L., Faloutsos, C., Identifying Web Browsing Trends and Patterns, Computer, IEEE Press, 2001, *7*

Montgomery02 Montgomery, A., Tenure of panel members in Montgomery Faloutsos Study, Personal Communication, 2002, *April*

Moxon96 Moxon, B., Defining Data Mining, DBMS, 1996, *9*, p. 11-13

Nagin93 Nagin, D. & Land, K., Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poission model, Criminology, 1993, *31*, p. 327-362

Nagin95 Nagin, D., Farrington, D. & Moffitt, T., Life-course trajectories of different types of offenders, Criminology, 1995, *33*, p. 111-139

Nagin99a Nagin, D., Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach, Psychological Methods, 1999, *2*, p. 139-157

Nagin99b Nagin, D., & Tremblay, R.E., Trajectories of boys' physical aggression opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency, Child Development, 1990, *70*, p. 1181-1196

Novak98a Novak, T. P., Hoffman, D. L., Bridging the Digital Divide: The Impact of Race on Computer Access and Internet Use, Working Paper, Owen Graduate School of Management, Vanderbuilt University, Nashville, TN, 1998

NTIA02 National Telecommunication and Information Administration, A Nation Online: How Americans Are Expanding Their Use of the Internet, Washington, D.C, Feb. 2002, http://www.ntia.doc.gov/ntiahome/dn/index.html [5-Nov-02]

NTIA95 http://www.ntia.doc.gov/ntiahome/fallingthru.html [5-Nov-02]

NTIA99b National Telecommunication and Informatino Administration, U.S. Department of Commerce, Falling Through the Net: Defining the Digital Divide". A Report on the Telecommunications and Information Technology Gap in America, July 1999, http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html [5-Nov-02]

NTIAa National Telecommunication and Informatino Administration, U.S. Department of Commerce, Falling Through the Net: Toward Digital Inclusion, http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html [5-Nov-02]

NUA NUA Internet: How many online?, http://www.nua.ie/surveys/how_many_online/world.html [5-Nov-02]

NUA01 Nua Internet Surveys, Web users start to turn off, Nielsen/NetRatings, 01-26-2001,

http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905356387&rel=true [5-Nov-02]

Oracle http://www.oracle.com/ [5-Nov-02]

Peppers97 Peppers, D., Rogers, M., Enterprise one to one: Tools for competition in the interactive age, New York, Doubleday, 1997

Perillieux00 Perillieux, R., Bernat, R., Bauer, M., Digitale Spaltung in Deutschland: Ausgangssituation, Internetionaler Vergleich, Handlungsempfehlungen - Booz-Allen & Hamilton Report, 2000

Perl http://www.perl.com/ [5-Nov-02]

Pine95 Pine, B. J., Peppers, D., Rogers, M., Do you want to keep your customers forever?, Havard Business Review, 1995, *March-April*, p. 103-114

Pool84 Pool, I., Inose, H., Takasaki, N., & Hurwitz, R., Communication flows: A census in the United States and Japan, New York, North-Holland, 1984

Puscher00 Puscher, F., Logfiles richtig lesen!, Internet World, 2000, *1*, p. 100-103,

Raftery95 Raftery, A. E., Bayesian model selection in social research, Sociological Methodology, 1995, *25*, p. 111-164

Reichheld00 Reichheld, F. F., Schefter, P., E-Loyalty - Your Secret Weapon on the Web, Harvard Business Review, 2000, *4*, p. 105-113

Reichheld90 Reichheld, F. F., Sasser, W. E. J., Zero Defections: Quality Comes to Services, Harvard Business Review, 1990, *5*, p .2-9

Rockman95 S. Rockman, In School or Out: Technology, Equity and the Future of our Kids, Communications of the ACM, 38, *6*, p. 25-29

Roeder99 Roeder, K., Lynch, K., & Nagin, D., Modeling uncertainty in latent class membership: A case study from criminology, Journal of the American Statistical Association, 1999, *33*, p. 766-777

SAN http://www.schulen-ans-netz.de [5-Nov-02]

Sarris90 Sarris, V., Methodologische Grundlagen der Experimentalpsychologie, Volume1, Munich, Reinhardt, 1990

Schaarschmidt01 Schaarschmidt, R., Nowitzky, J. L., Clickstream Warehousing für e-CRM: Neue Herausforderungen an die Datenhaltung?, Buhl, H. U.,

Huther, A., Reitwiesner, B.: Information Age Economy. 5. Internationale Tagung Wirtschaftsinformatik, Heidelberg, Physica-Verlag, 2001, p. 117-131

Schafer99 Schafer, J. B., Konstan, J., Riedl, J., Recommender Systems in E-Commerce, Proceedings of the ACM Conference on Electronic Commerce, Denver, CO, 1999, p. 158-166

Schmitt99 Schmitt, E., Manning, H., Paul, Y., Tong, J., Measuring Wen Success, Forrester Report, 1999, *11*

Scholz01 Scholz, J., Moderates Wachstum, e-Market, 2001, *47/48*, p. 53-54,

Schwickert01 Schwickert, A.C., Wendt, P., Web-Logfile-Analyse, Praxis der Wirtschaftsinformatik, 2001, *221*, p. 95-103

Shade Shade, L. R., Using A Gender-based Analysis in Developing a Canadian Access Strategy: Backgrounder Report - Universal Access Project, 2002, http://www.fis.utoronto.ca/research/iprp/ua/gender/GenderBased.html [5-Nov-02]

SourceForge http://sourceforge.net/projects/analog [5-Nov-02]

Spiekermann00 Spiekermann, S., Christ, M., Designing Recommender Systems Strategically, Proc. 3rd Berlin Internet Economics Workshop, Berlin, 2000

Staudt01 Staudt, E., Die mobile Gesellschaft, Buhl, H. U., Huther, A., Reitwiesner, B.: Information Age Economy. 5. Internationale Tagung Wirtschaftsinformatik, Heidelberg, Physica-Verlag, 2001, p. 15-28

Tauscher97a Tauscher, L., Greenberg, S., How people revisit web pages: empirical findings and implications for the design of history systems, International Journal of Human Computer Studies, Special Issue on World Wide Web Usability, 1997, *47*, p. 97-138

Teltzrow01 Teltzrow, M., Günther, O., eCRM:Konzeption und Möglichkeiten zur Effiziensmessung, Praxis der Wirtschaftsinformatik, 2001, *221*, p. 16-26

Titterington85 Titterington, D.M., Smith, A.F.M. & Makov, U.E., Statistical analysis of finite mixture distributions, New York, Wiley, 1995

TRAJ SAS Proc TRAJ, http://www.stat.cmu.edu/~bjones [5-Nov-02]

TRUSTe http://www.truste.org [5-Nov-02]

USCensus00 U.S. Census Bureau, USA Statistics in Brief, http://www.census.gov/statab/www/brief.html [5-Nov-02]

Webgauge http://www.gfk-webgauge.com [5-Nov-02]

Webtrends http://www.netiq.com/solutions/analytics/default.asp [5-Nov-02]

Wiggins01 Wiggins, R. W., The Effects of September 11 on the Leading Search Engine, First Monday, 2001, *10*

Wilde99 Wilde, E., World Wide Web, Berlin, Springer, 1999, p. 110

Wittig00 Wittig, P., Die datenschutzrechtliche Problematik der Anfertigung von Persönlichkeitsprofilen zu Marketingzwecken, Recht der Datenverarbeitung, 2000, *2*, p. 59-62

WSJ01 Ads Click, 2001, http://www.wsj.com/articles/SB1004115312686358960.htm [5-Nov-02]

Yahoo99a Product Analysis of Yahoo!, http://market.econ.Vanderbilt.edu/ba250/spring99/ Yahoo/product.html [5-Nov-02]

Yoo01 Yoo, B., Naveen, D., Developing a Scale to Measure the Perceived Quality of an Internet Shopping Site (SITEQUAL), Quarterly Journal of Electronic Commerce, 2001, *1*, p. 31-36

Zeithaml96 Zeithaml, V. A., Berry. L. L., Parasuraman, A., The Behavioral Consequences of Service Quality, Journal of Marketing, 1996, *April*, p. 31-46

Appendix A – Further Statistics about the HomeNet Project at Carnegie Mellon University

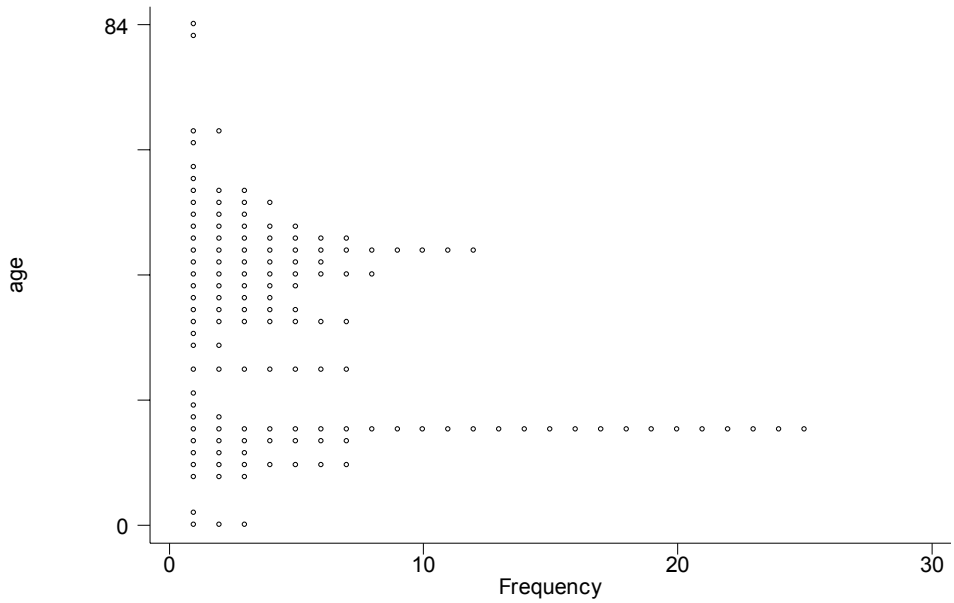


Figure 48: Age distribution of individuals in the HomeNet sample

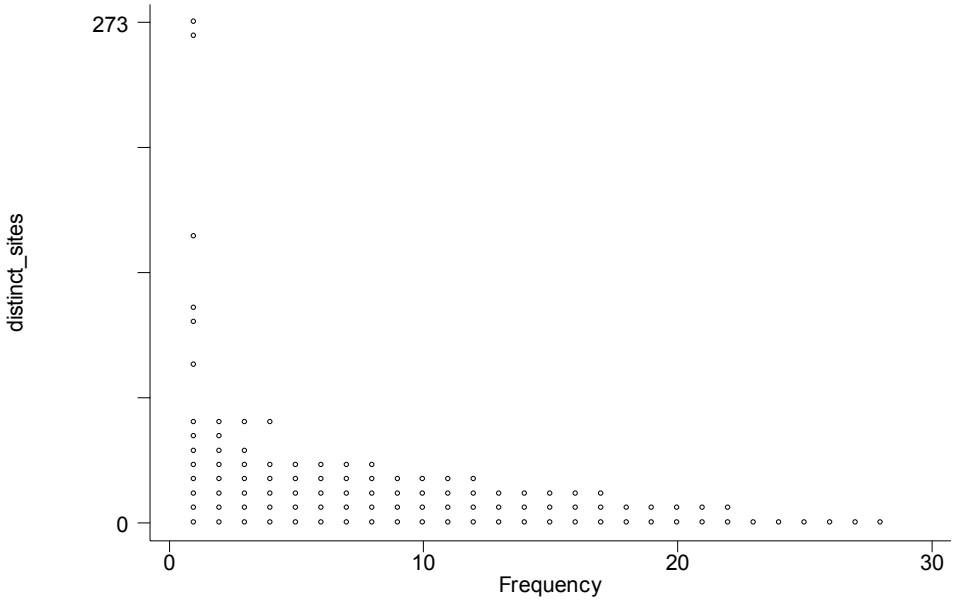


Figure 49: Distribution of the number of unique Web sites visited monthly by individuals in the HomeNet sample

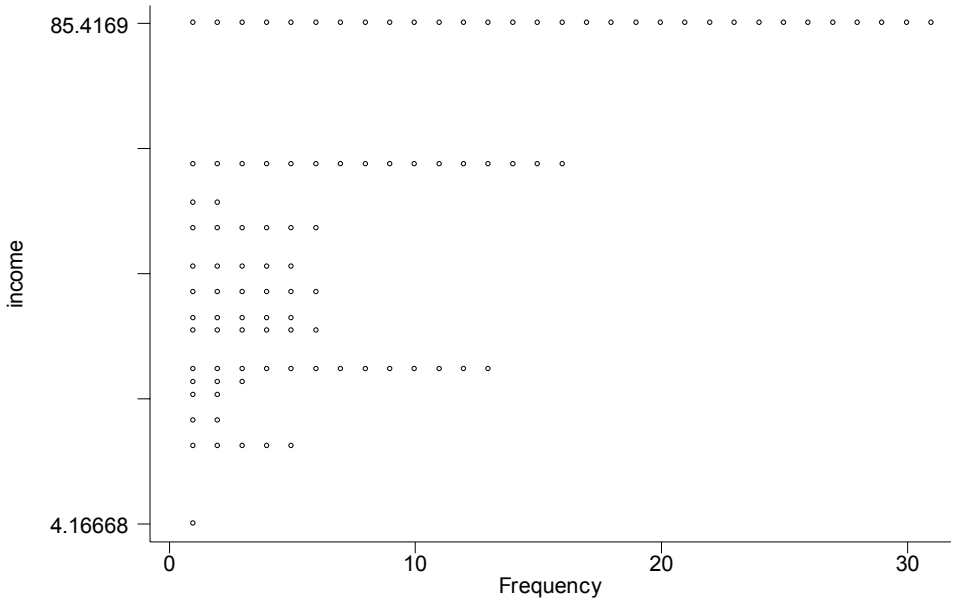


Figure 50: Distribution of household income in the HomeNet sample

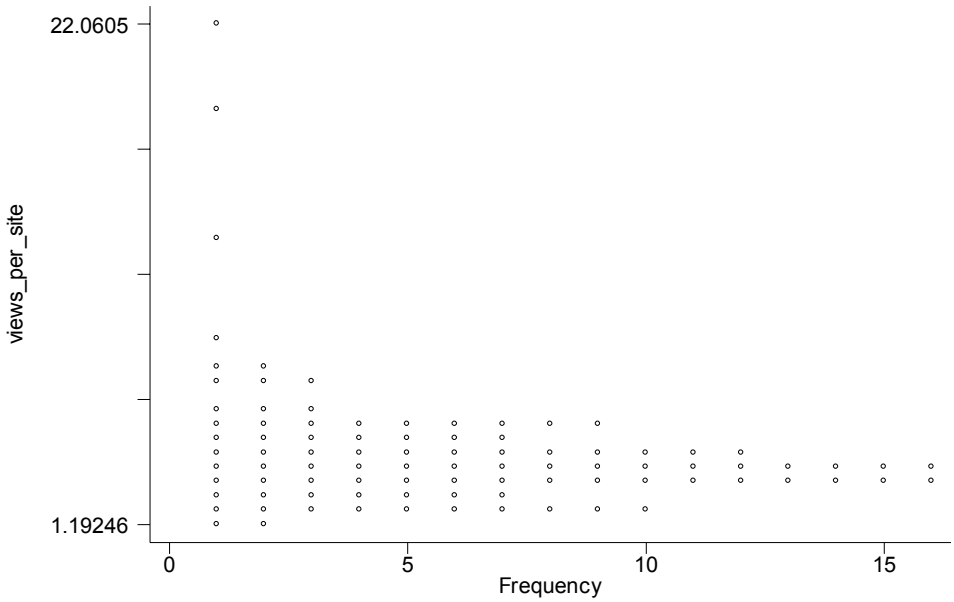


Figure 51: Distribution of pages viewed by individuals monthly in the HomeNet sample

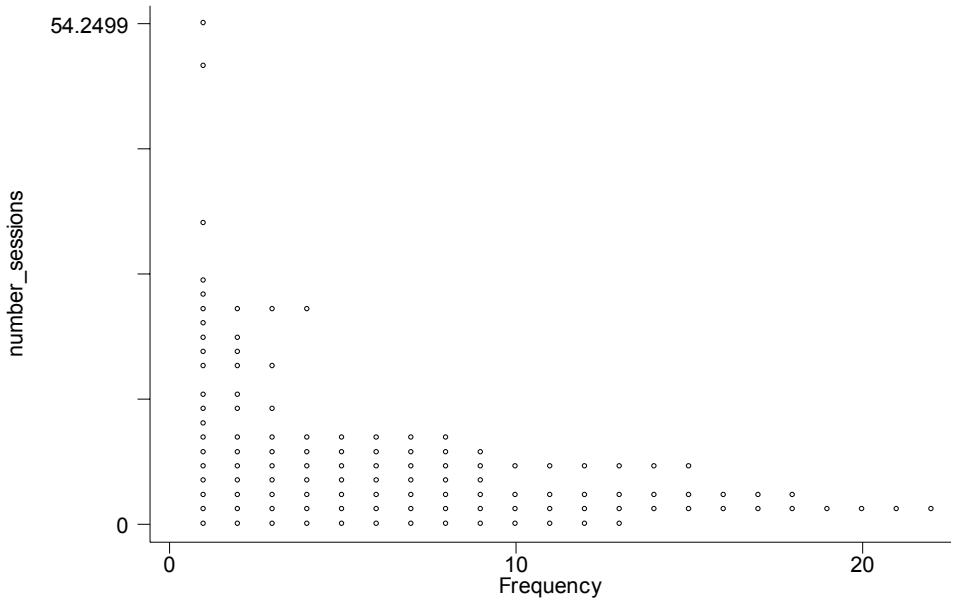


Figure 52: Distribution of the individual number of Web sessions in the HomeNet sample

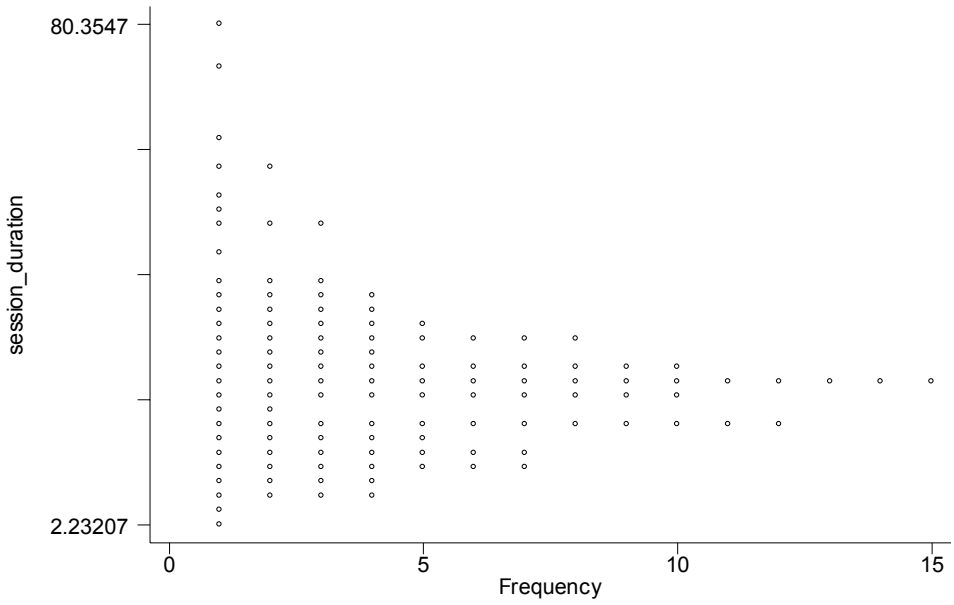


Figure 53: Distribution of duration of Web sessions in the HomeNet sample

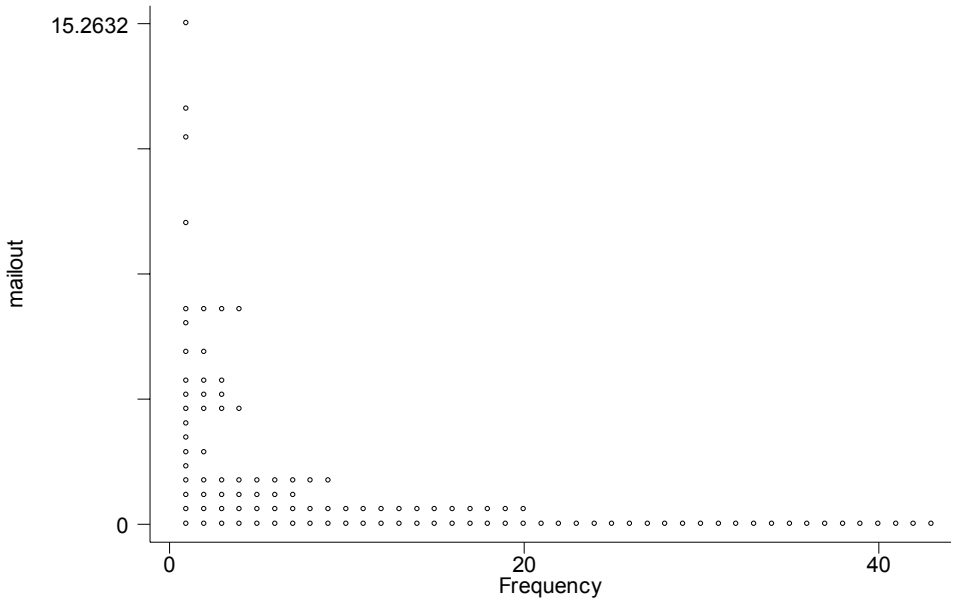


Figure 54: Distribution of number of emails sent weekly by individuals in the HomeNet sample

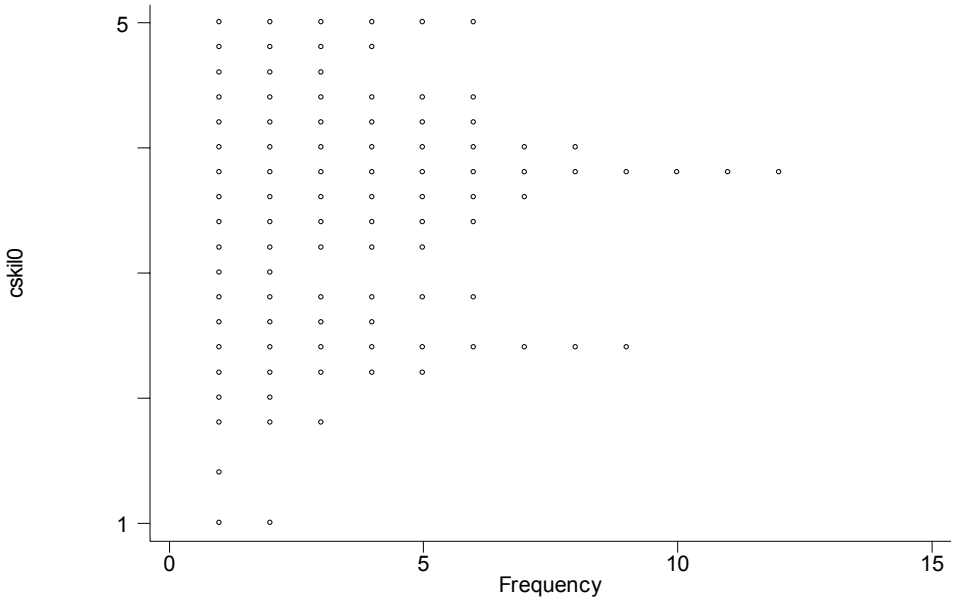


Figure 55: Distribution of computer skill level in the HomeNet sample

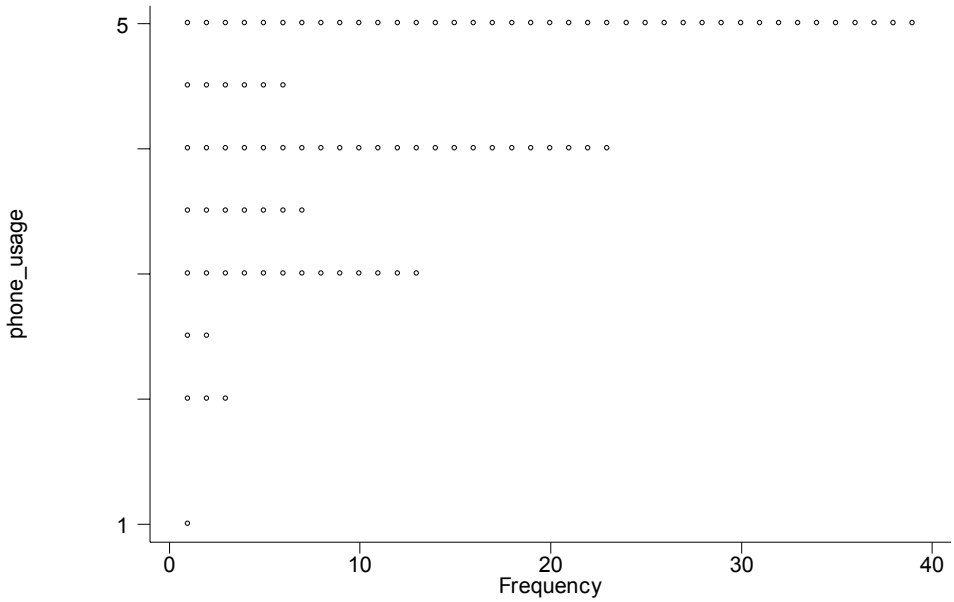


Figure 56: Distribution of phone usage in the HomeNet sample

Tab. 28: Group profiles of HomeNet users

	All users	low-rate users	Moderate users	Heavy users	very heavy users	significance
Percentage	100%	52.5%	30.2%	13.7%	3.6%	
Adult	74.3%	70.9%	77.4%	78.6%	80.0%	Prob > chi2 = 0.8700 (logit)
Female	51.4%	56.4%	58.1%	28.6%	20.0%	Prob > chi2 = 0.1022 (logit)
Minority	27.6%	32.7%	29.0%	7.1%	20.0%	Prob > chi2 = 0.1992 (logit)
Position:						
Mom	25.9%	26.0%	28.6%	21.1%	20.0%	Prob > chi2 = 0.9205 (logit)
Dad	20.1%	16.4%	21.4%	31.6%	20.0%	Prob > chi2 = 0.5571 (logit)
Daughter	20.1%	21.9%	21.4%	15.8%	0.0%	Prob > chi2 = 0.8298 (logit)
Son	17.3%	17.8%	14.3%	21.1%	20.0%	Prob > chi2 = 0.9200 (logit)
Other	16.5%	17.8%	14.3%	10.5%	40.0%	Prob > chi2 = 0.5022 (logit)
Avg. age	31.91	32.22	32.76	28.47	33.40	Prob > F = 0.8327 (regress)
Community involvement	3.70	3.46	3.77	4.22	4.24	Prob > F = 0.1343 (regress)
Computer skills	3.43	3.46	3.11	3.90	3.84	Prob > F = 0.0725 (regress)
Innovativeness	3.32	3.20	3.45	3.44	3.40	Prob > F = 0.1192 (regress)
Connection hrs weekly	2.15	1.3	2.25	2.72	9.26	Prob > F = 0.0000 (regress)
Mail sent weekly	1.62	1.27	1.75	1.98	3.76	Prob > F = 0.2509 (regress)
Household income	54.41	55.92	52.77	56.67	42.50	Prob > F = 0.6377 (regress)

Appendix B – The digital divide (additional figures)

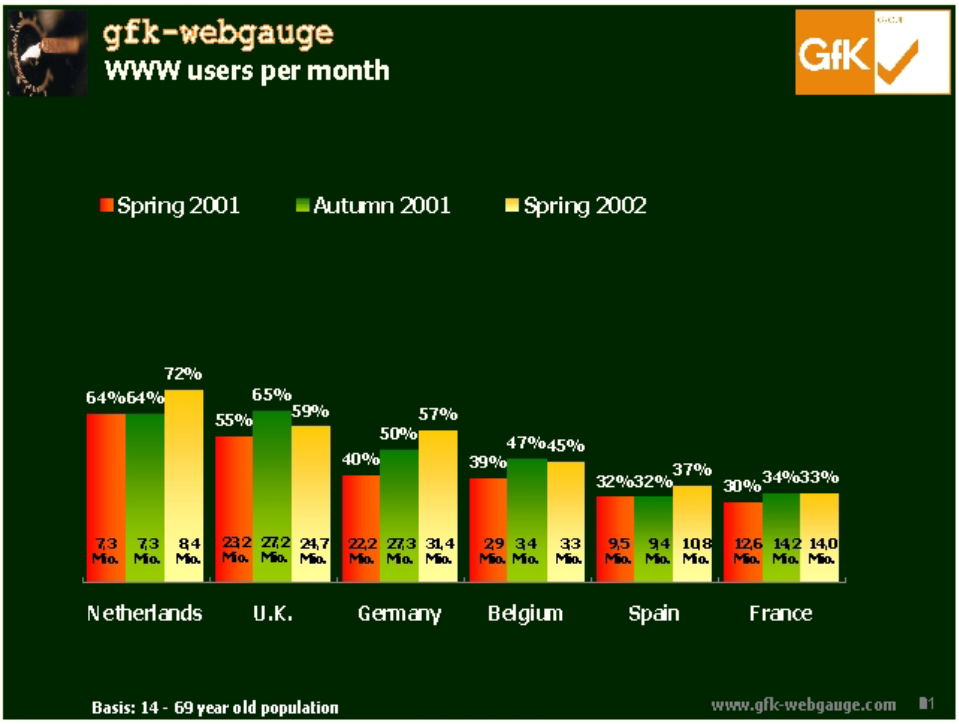


Figure 57: Web users per month in European countries [Webgauge]

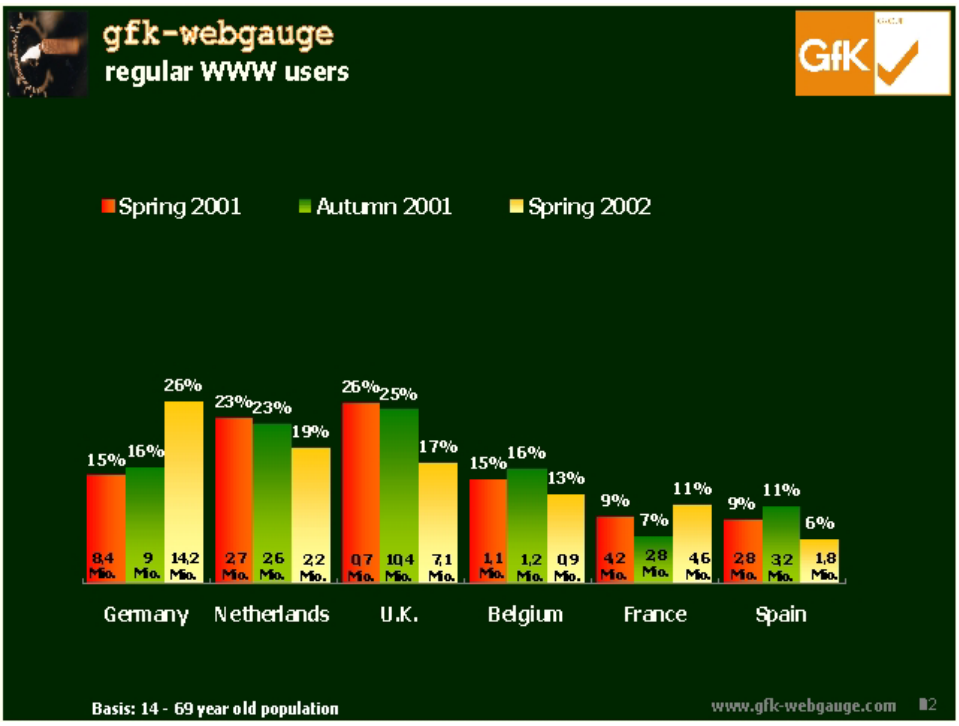


Figure 58: Regular Web users in European countries [Webgauge]

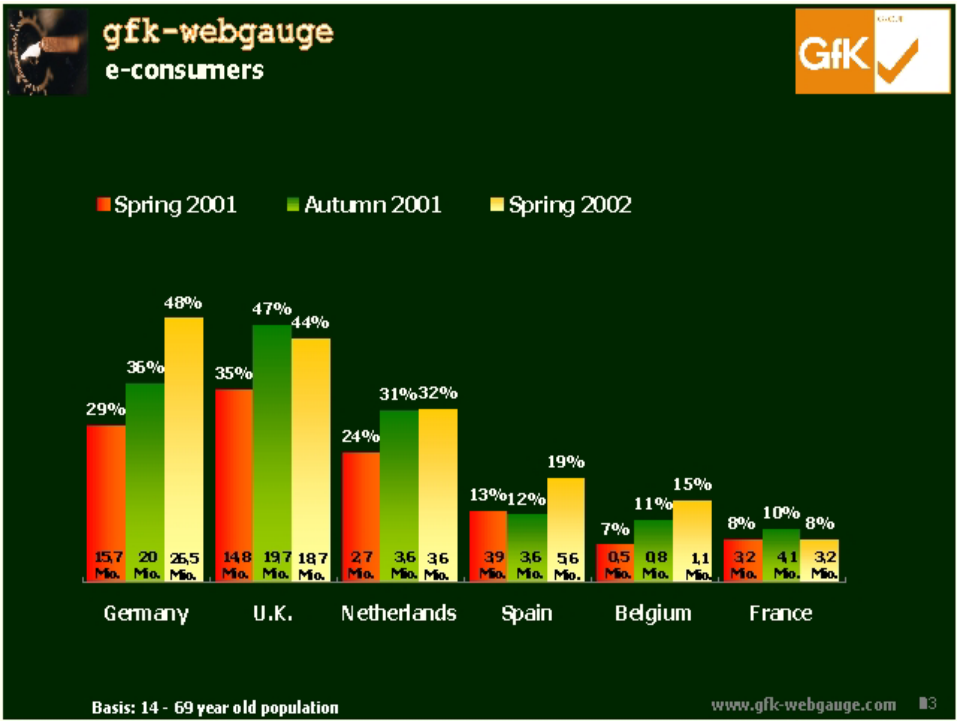


Figure 59: e-consumers in European countries [Webgauge]

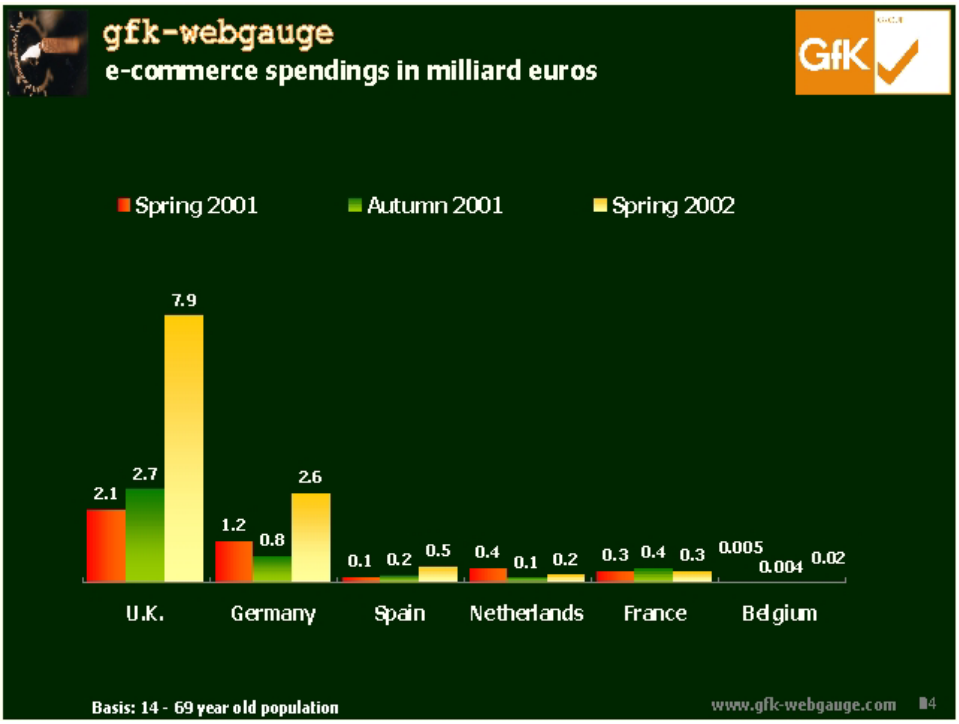


Figure 60: e-commerce spendings in European countries [Webgauge]

Figure 61: Infrastructure: Hard and soft factors in Germany [Perillieux00]

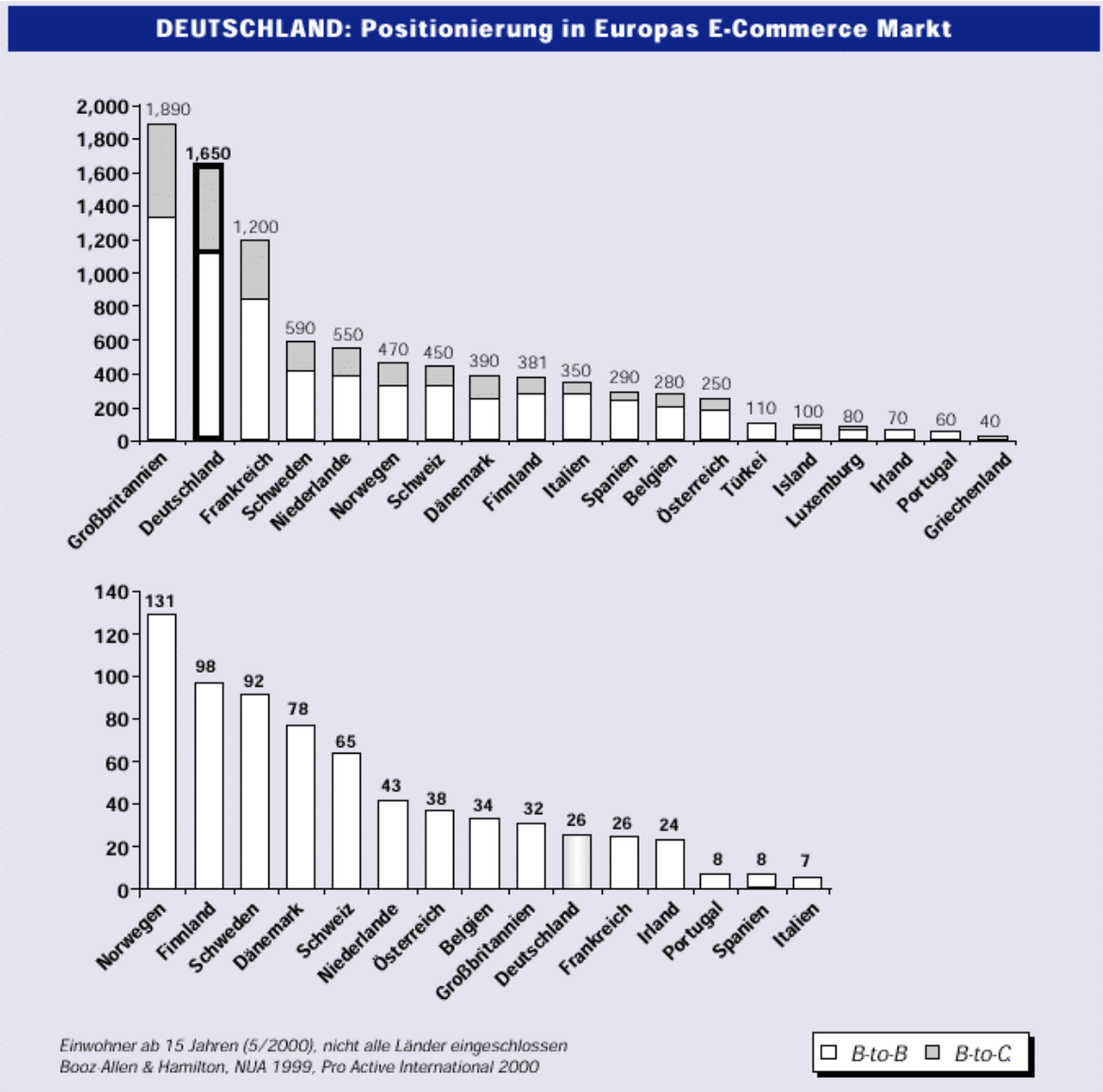


Figure 62: Europe's e-commerce position [Perillieux00]

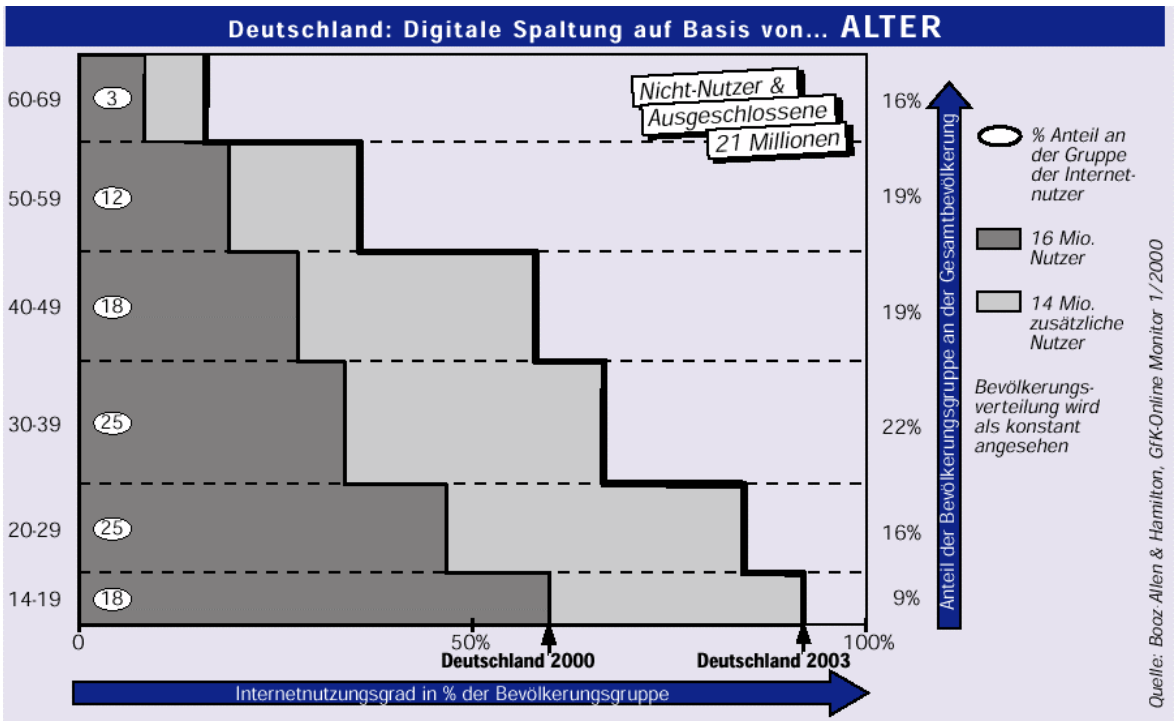


Figure 63: Age determines Internet access in Germany [Perillieux00]

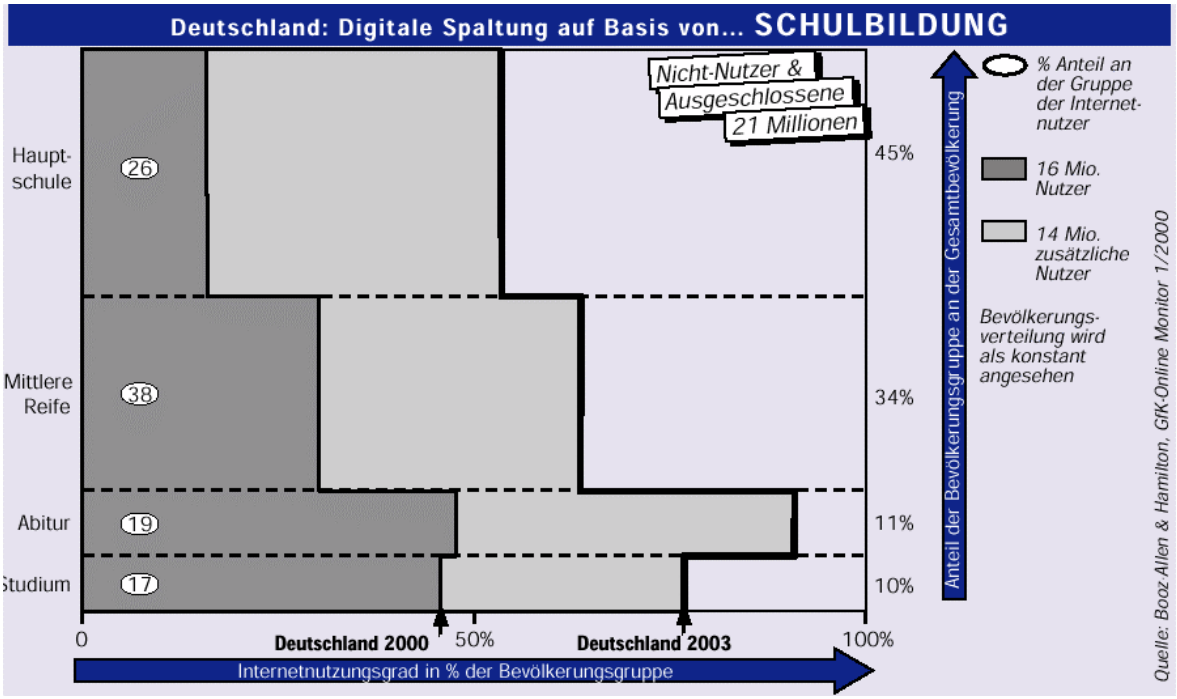


Figure 64: Education determines Internet access in Germany [Perillieux00]

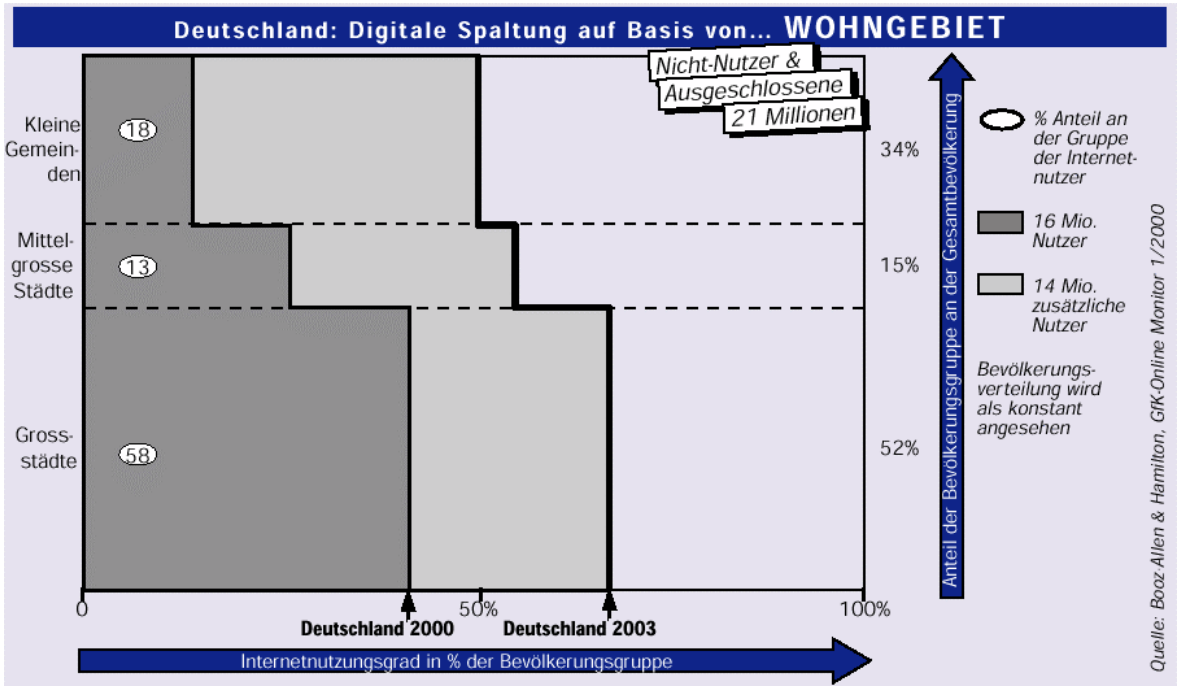


Figure 65: Region determines Internet access in Germany [Perillieux00]

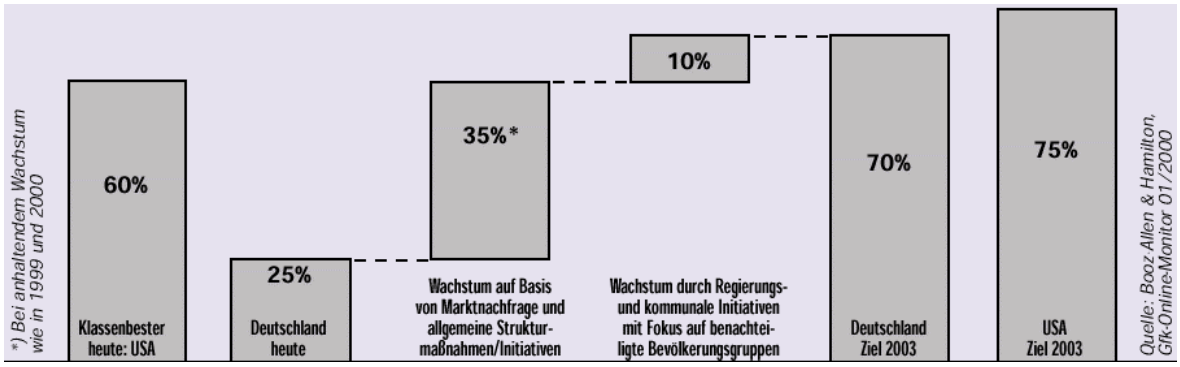


Figure 66: Internet penetration in Germany and the United States [Perillieux00]

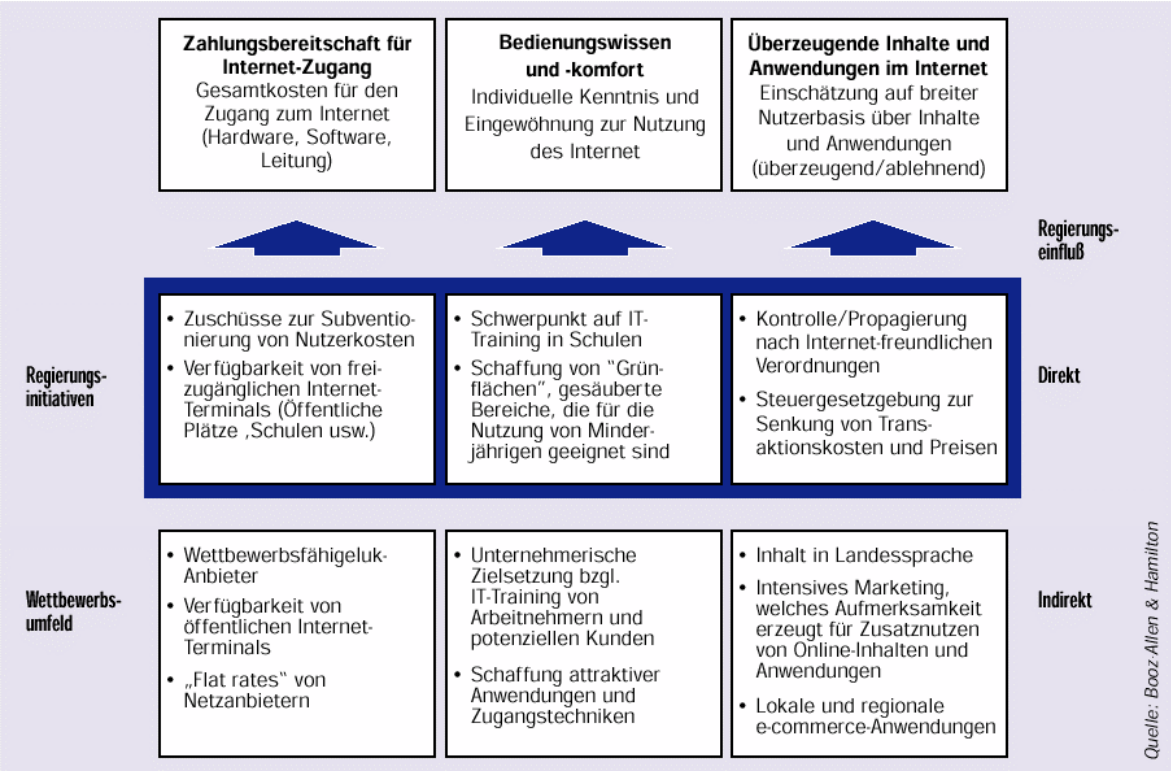


Figure 67: Drivers and impediments of Internet usage [Perillieux00]

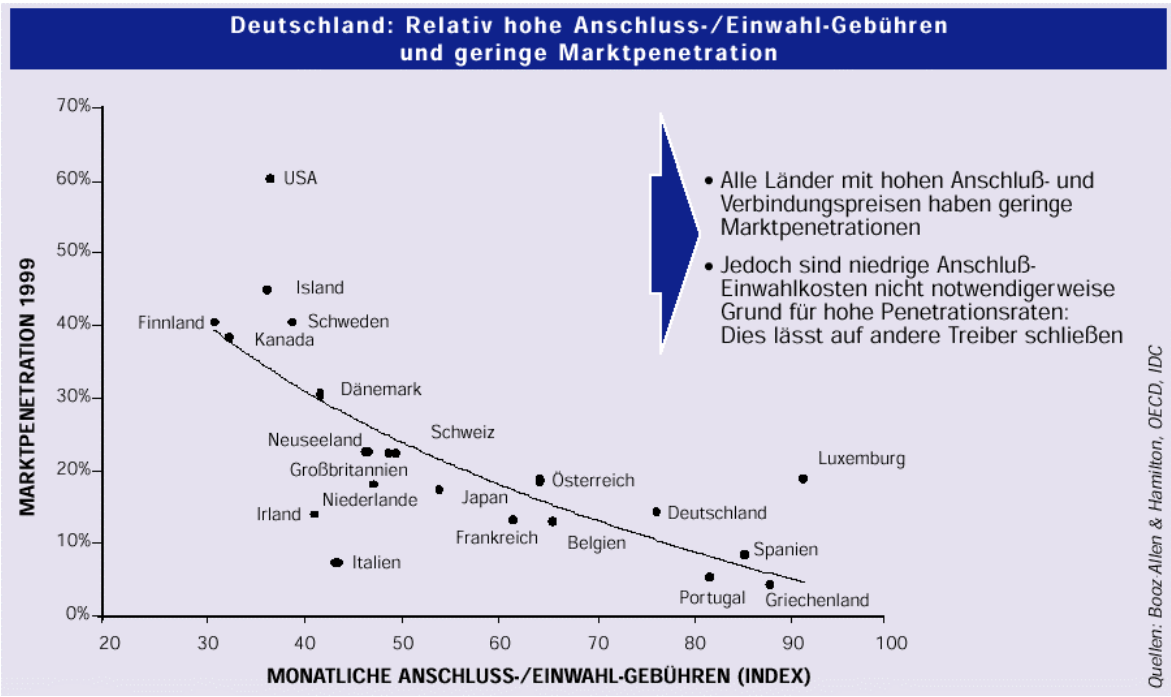


Figure 68: Telecommunication cost determine Internet penetration

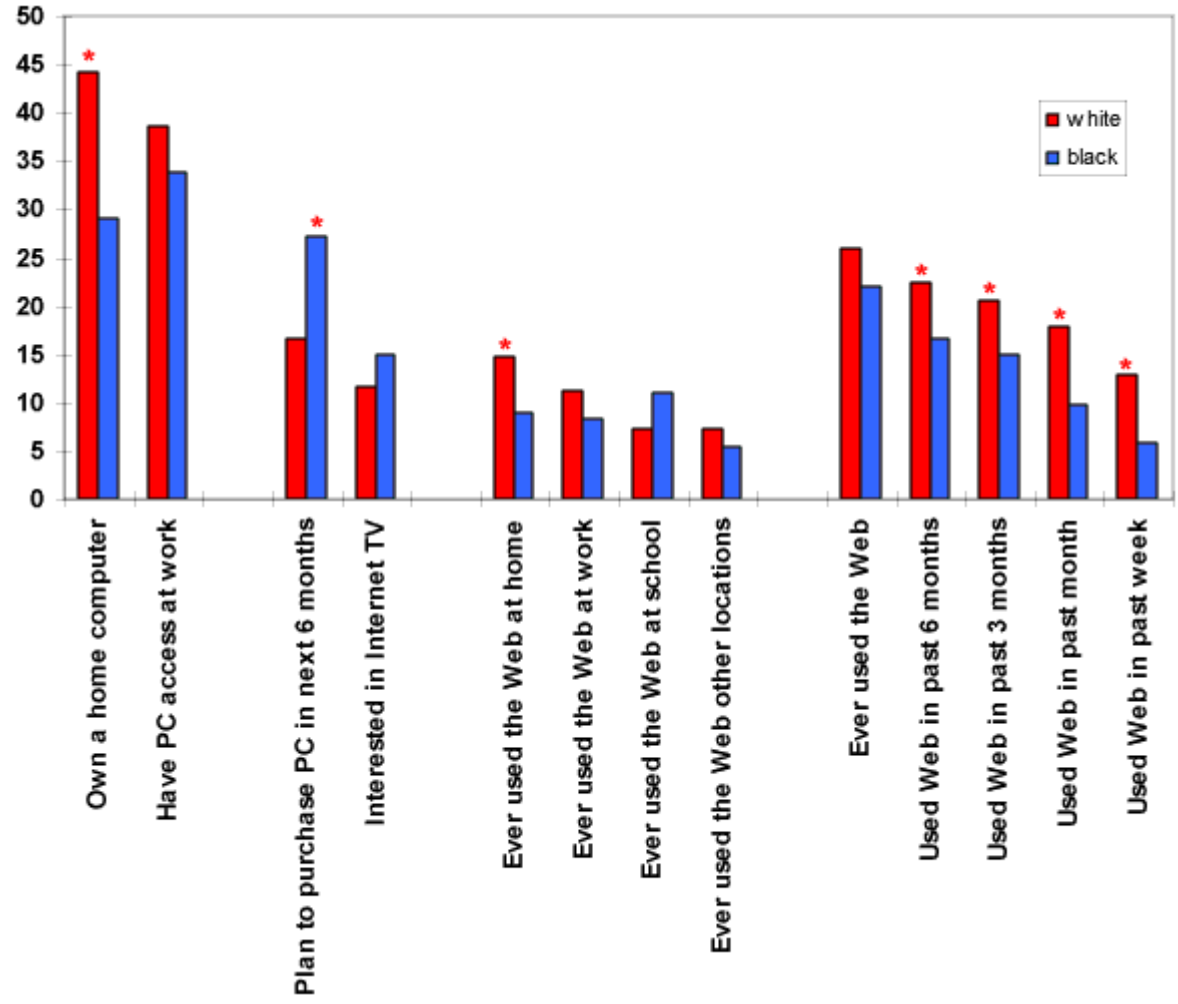


Figure 69: Race determines computer access at work and at home, and Internet usage [Novak98a]

„**Note:** Asterisk (*) indicates p<.05. All significance tests were obtained using Research Triangle Institute's SUDAAN software and incorporating sampling weights. Sampling weights provided by Nielsen Media Research specified the probability of a respondent being selected into the sample. These sampling weights were adjusted for number of phones in the household and number of people aged 16 and older in the household, and were also adjusted for nonresponse by post-stratification adjustments to equate sample race, education, age, and gender distributions to Census data (Nielsen Media Research 1997)"

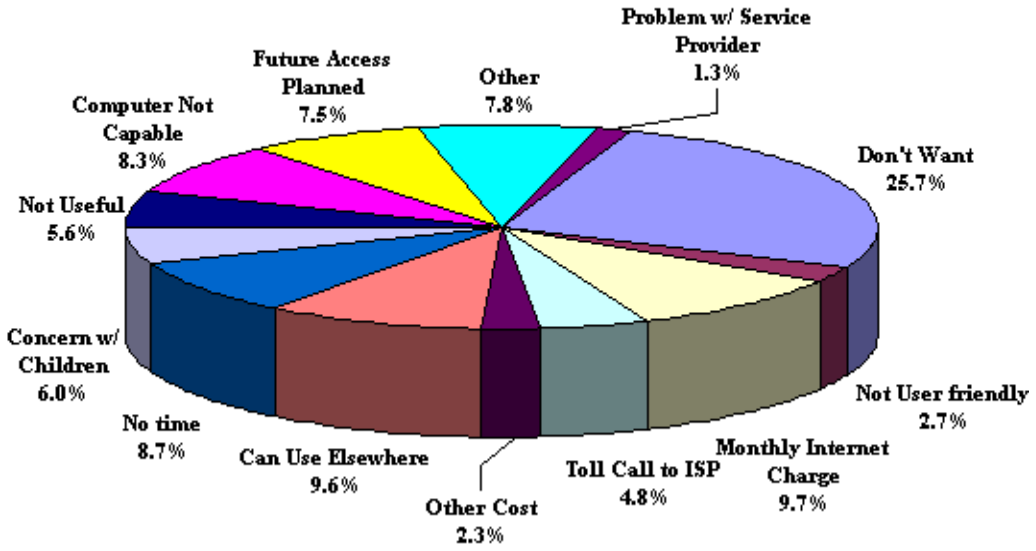


Figure 70: Reasons for United States' households with a computer/webTV® not using the Internet at home [NTIA99b]

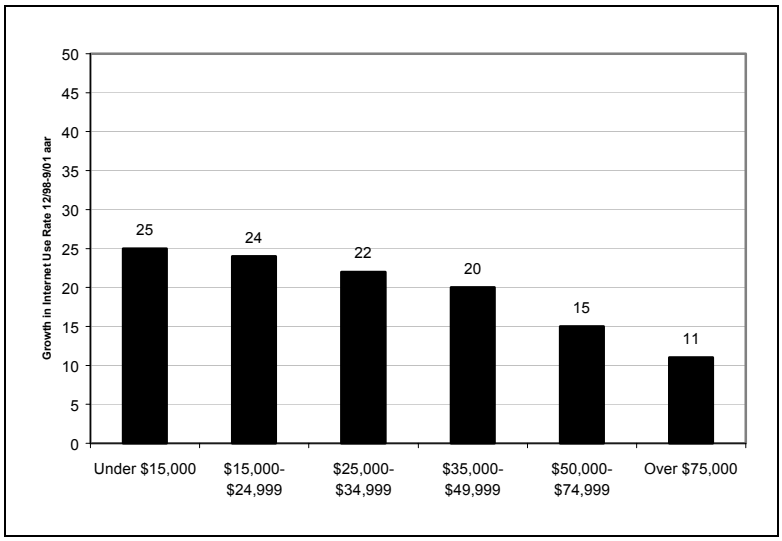


Figure 71: Growth rate in Internet use by family income (annual rate December 1998 to September 2001)

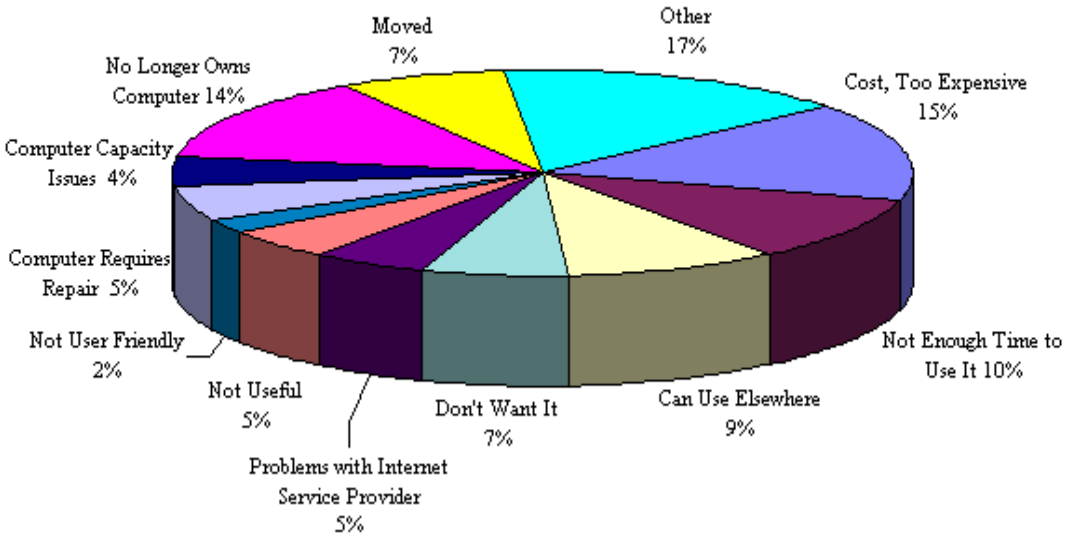


Figure 72: Reasons for United States' households discontinuing Internet use [NTIA99b]

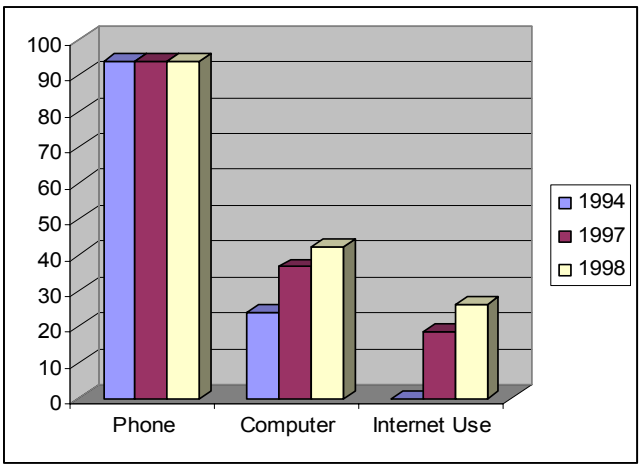


Figure 73: Percent of United States' households with a telephone, computer, and Internet use

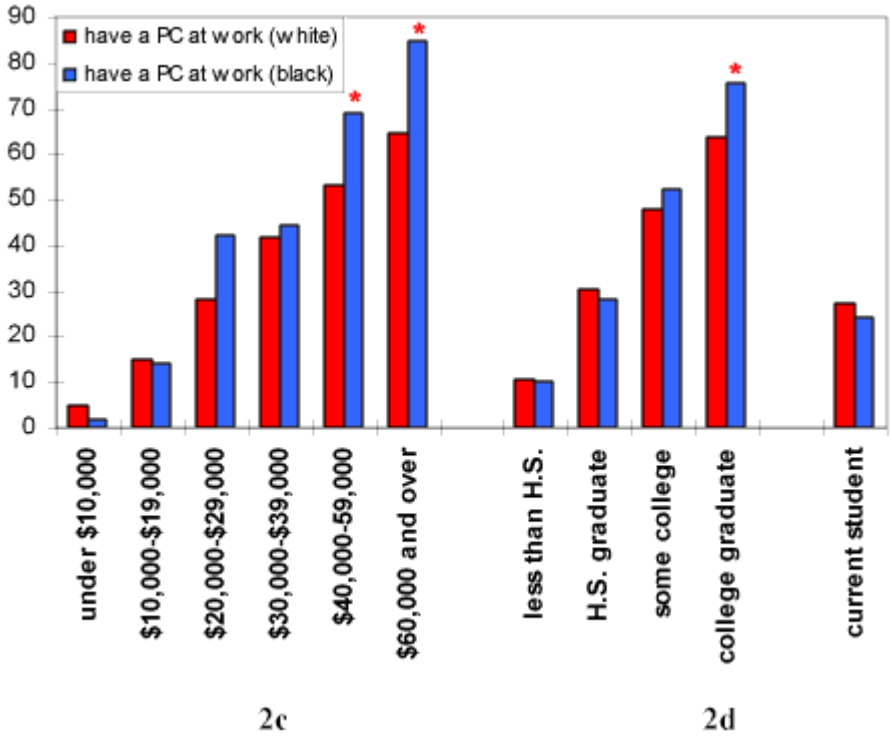


Figure 74: Income and education determines PC at work [Novak98a]

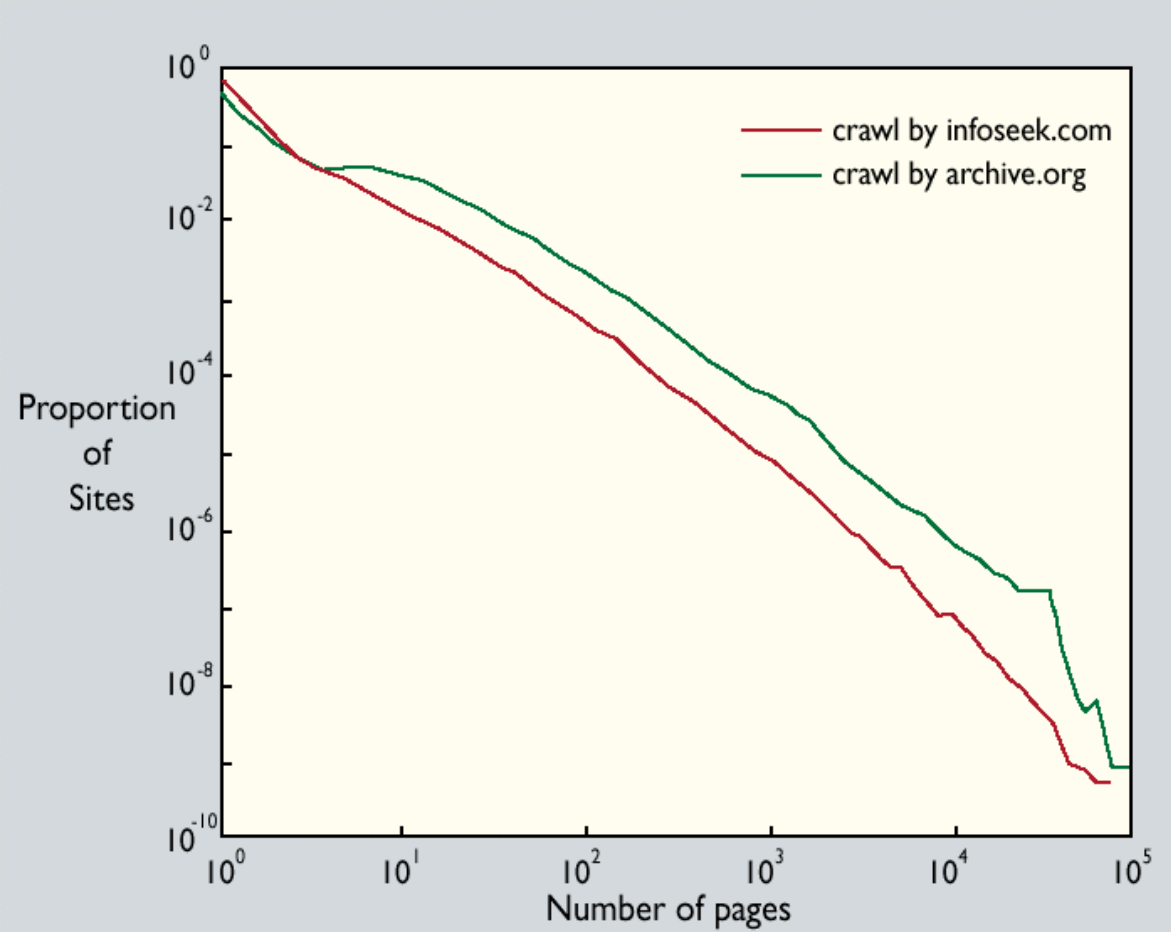


Figure 75: Proportion of sites and number of pages [Adamic01]

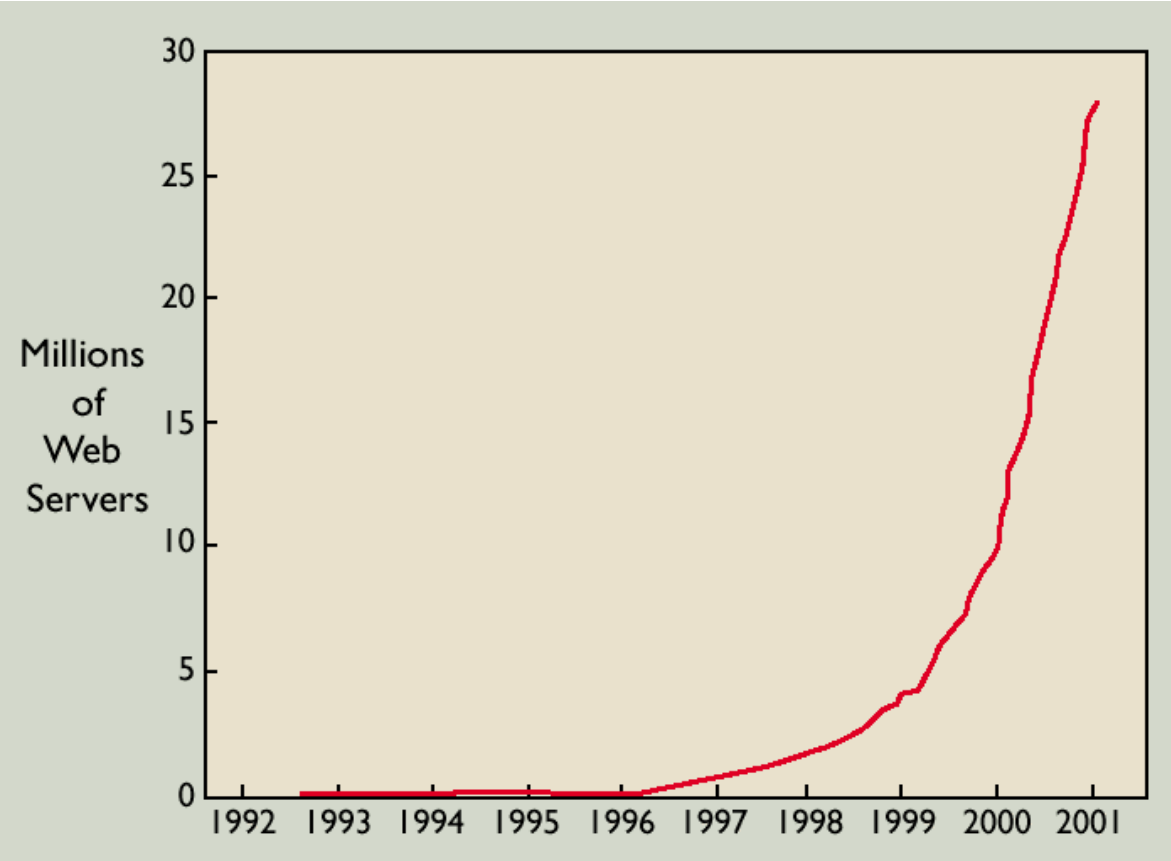


Figure 76: The exponential growth of the Internet [Adamic01]

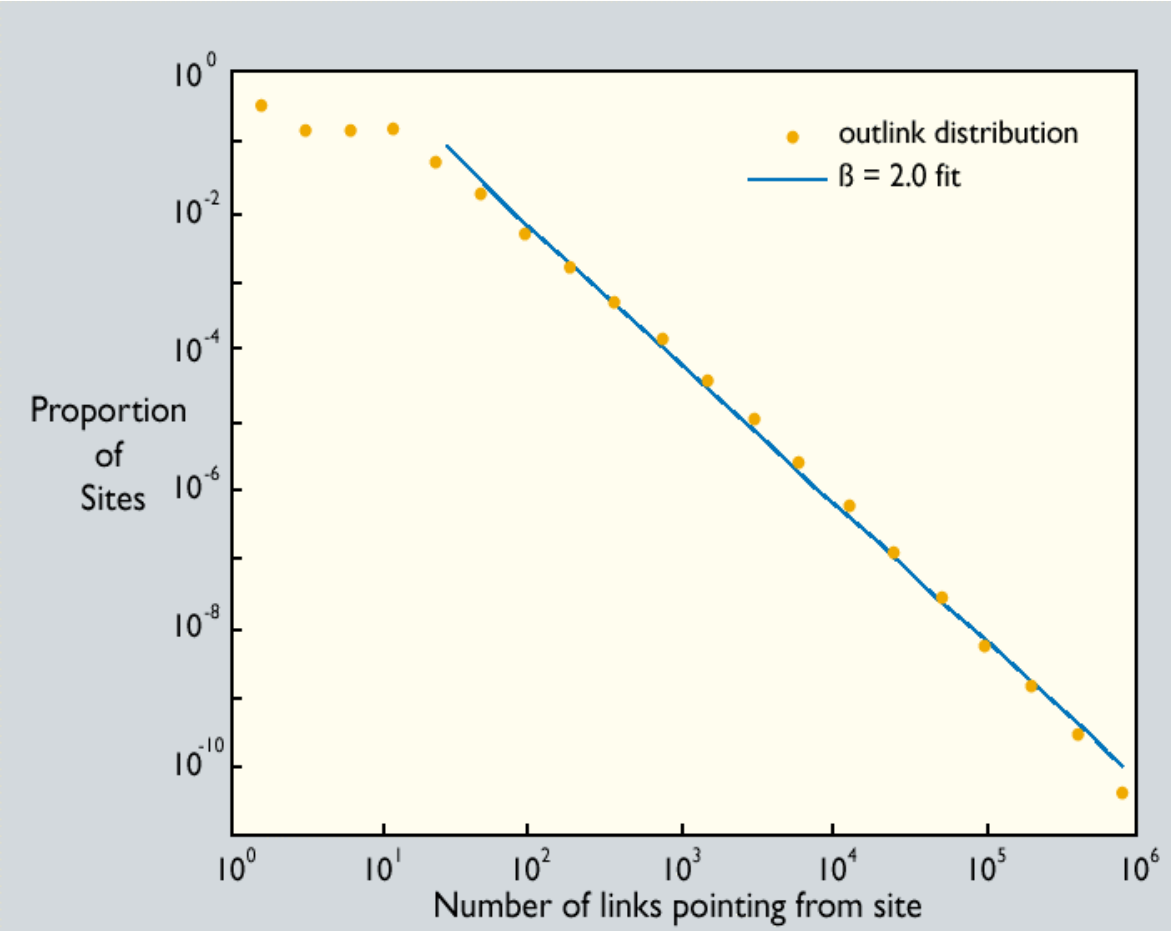


Figure 77: Proportion of sites and number of links pointing from site [Adamic01]

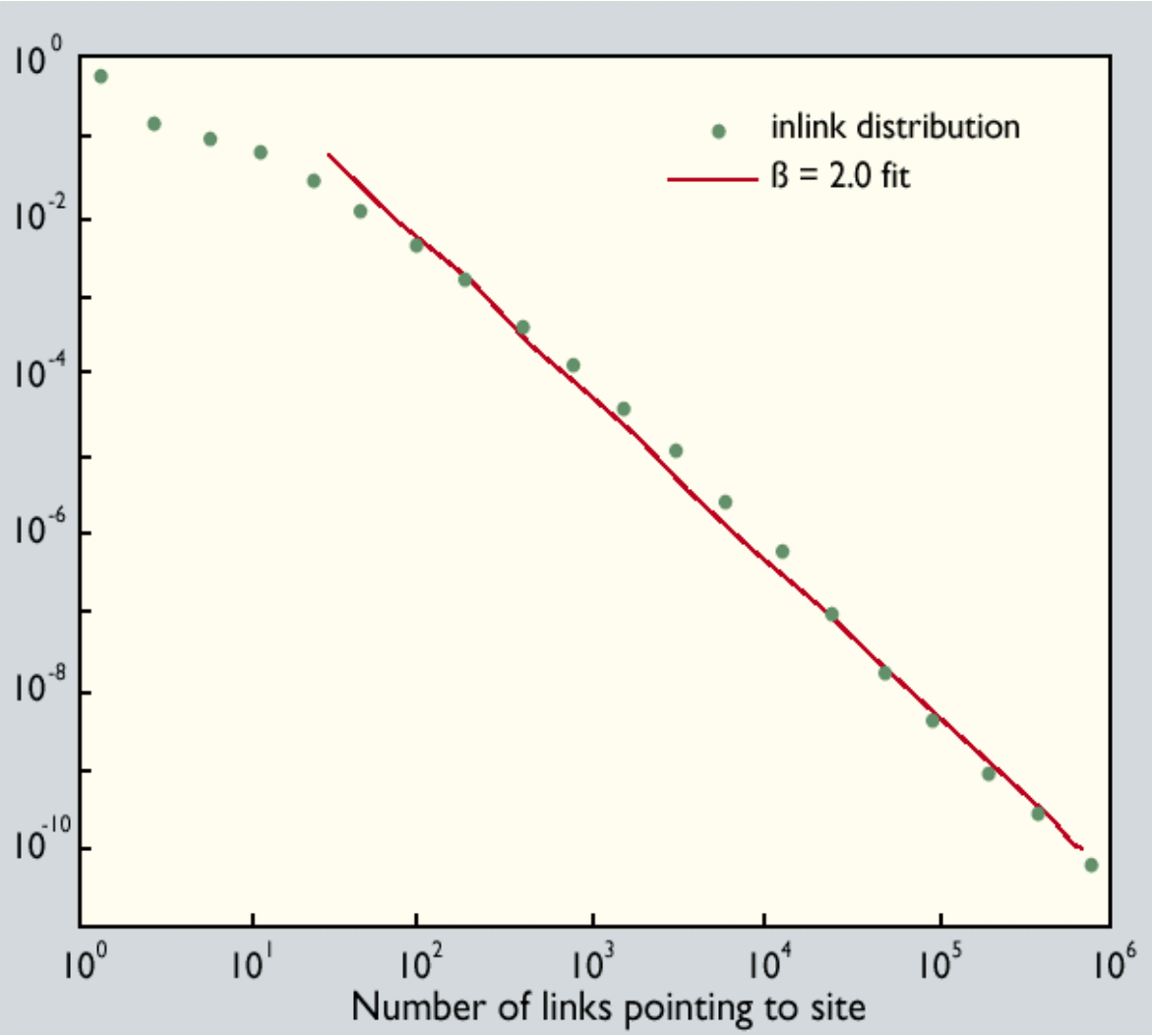


Figure 78: Proportion of sites and number of links pointing to site [Adamic01]

Acknowledgements

Writing this dissertation would not have been possible without the support of several mentors. I would like to thank Prof. Oliver Guenther, Prof. Ramayya Krishnan, and Prof. Daniel S. Nagin for their guidance, feedback, and comments. In particular, I would like to thank Daniel Nagin for providing the statistical method on which the results on saturation of lay Web usage are based on.

Further, I would like to thank the institutions that provided the necessary funding for this research:

HomeNet is funded by grants from the National Science Foundation under Grant No. IRI-9408271, Apple Computer, AT&T, Bell Atlantic, Bellcore, Intel, Carnegie Mellon University's Information Networking Institute, Interval, the Markle Foundation, the NPD Group, the United States Postal Service, and US West. Farallon Computing and Netscape Communications contributed software

Development of the trajectory estimation method and software was supported by the National Science Foundation under Grant No. SBR-9513040 to the National Consortium on Violence and also by separate National Science Foundation grants SBR-9511412 and SES-9911370.

The author Mario Christ was supported by the German Research Society, Berlin-Brandenburg, Graduate School in Distributed Information Systems (DFG grant no. GRK~316). This research was also supported by the TransCoop program of the Alexander von Humboldt Foundation, Bonn, Germany.

The work of Ramayya Krishnan was funded in part by NSF grant CISE/IIS/KDI 9873005.

The work of Robert Kraut was funded in part by NSF grant IIS-9980013,

Empfangene Unterstützung und Hilfe durch Kollegen:

- Prof. Daniel S. Nagin lieferte die methodologische Grundlage für die Arbeit in Kapitel 3. Von ihm stammt die statistische Methode zur „Trajectory Analysis". Ferner wurden alle Schritte in Kapitel 3, 5, und 6 mit ihm inhaltlich diskutiert.
- Prof. Ramayya Krishnan lieferte wertvolle Ideen und Anregungen fuer die Kapitel 3, 5, und 6. Die Arbeit in diesen Kapiteln wurde erst nach inhaltlicher Diskussion und mehreren Revisionen durch ihn auf Konferenzen publiziert.
- Prof. Oliver Günther ist als Doktorvater der Dissertation direkt an der Grundausrichtung der Arbeit beteiligt. Sämtliche Kapitel durchliefen einen Review-Prozess seinerseits.
- Stefan Baron half beim „Sessionizing" der HomeNet Logfiles in Kapitel 4.
- In Kapitel 3 sind Kommentare eines Review Prozesses durch Gutachter der 34th Hawaii International Conference on System Science eingeflossen.
- In Kapitel 4 sind Kommentare eines Review Prozesses durch Gutachter der 36th Hawaii International Conference on System Science eingeflossen.
- In Kapitel 5 sind Kommentare eines Review Prozesses durch Gutachter der 10th European Conference on Information Systems eingeflossen.
- In Kapitel 6 sind Kommentare eines Review Prozesses durch Gutachter der 35th Hawaii International Conference on System Science eingeflossen.
- B. L. Jones implementierte die statistische Methode zur „Trajectory Analysis" als „SAS Proc", welches in Kapitel 3 diskutiert wird.

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Berlin, 29. November 2002

Mario Christ

Eidesstattliche Erklärung

Hiermit erkläre ich, Mario Christ, dass ich mich bisher noch an keiner Institution einem Doktorexamen unterzogen habe. Ferner wurde die Dissertation bisher noch an keiner anderen Fakultät vorgelegt.

Berlin, den 29. November 2002

Mario Christ